TUGAS AKHIR – SS 141501

PREDIKSI CURAH HUJAN MELALUI *MODEL OUTPUT STATISTICS* MENGGUNAKAN *CLASSIFICATION AND REGRESSION TREES* DENGAN *PRE-PROCESSING PRINCIPAL COMPONENT ANALYSIS*

ULUL AZMI
NRP 1311 100 702

Dosen Pembimbing
Dr. Sutikno, S.Si, M.Si

PROGRAM STUDI S1
JURUSAN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2017

TUGAS AKHIR – SS 141501

PREDIKSI CURAH HUJAN MELALUI *MODEL OUTPUT STATISTICS* MENGGUNAKAN *CLASSIFICATION AND REGRESSION TREES* DENGAN *PRE-PROCESSING PRINCIPAL COMPONENT ANALYSIS*

ULUL AZMI
NRP 1311 100 702

Dosen Pembimbing
Dr. Sutikno, S.Si, M.Si

PROGRAM STUDI S1
JURUSAN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2017

FINAL PROJECT – SS 141501

PREDICTION OF RAINFALL BY MODEL OUTPUT STATISTICS USING CLASSIFICATION AND REGRESSION TREES WITH PRE-PROCESSING PRINCIPAL COMPONENT ANALYSIS

ULUL AZMI
NRP 1311 100 702

Supervisor
Dr. Sutikno, S.Si, M.Si

UNDERGRADUATE PROGRAMME
STATISTICS DEPARTEMENT
FACULTY OF MATHEMATICS AND NATURAL SCIENCES
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2017

LEMBAR PENGESAHAN

PREDIKSI CURAH HUJAN MELALUI *MODEL OUTPUT STATISTICS* MENGGUNAKAN *CLASSIFICATION AND REGRESSION TREES* DENGAN *PRE-PROCESSING PRINCIPAL COMPONENT ANALYSIS*

TUGAS AKHIR
Diajukan untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Sains
pada
Program Studi S1 Jurusan Statistika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Teknologi Sepuluh Nopember

Oleh:
ULUL AZMI
NRP 1311 100 702

Disetujui oleh Dosen Pembimbing Tugas Akhir:

Dr. Sutikno, S.Si, M.Si
NIP : 19710313 199702 1 001 ( )

Mengetahui
Ketua Jurusan Statistika FMIPA-ITS

Dr. Suhartono
NIP. 19710929 199512 1 001

SURABAYA, JANUARI 2017

(Halaman ini sengaja dikosongkan)

ABSTRAK

PREDIKSI CURAH HUJAN MELALUI *MODEL OUTPUT STATISTICS* MENGGUNAKAN *CLASSIFICATION AND REGRESSION TREES* DENGAN *PRE-PROCESSING PRINCIPAL COMPONENT ANALYSIS*

Nama Mahasiswa : Ulul Azmi
NRP : 1311 100 702
Jurusan : Statistika
Dosen Pembimbing : Dr. Sutikno, S.Si, M.Si

ABSTRAK

Kondisi cuaca di Indonesia diumumkan untuk jangka waktu sekitar 24 jam melalui prakiraan cuaca hasil analisis Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). Sejak tahun 2004, BMKG telah berupaya melakukan penelitian untuk prakiraan cuaca jangka pendek dengan menggunakan data komponen cuaca *Numerical Weather Prediction* (NWP). Namun *output NWP* masih sering bias, sehingga perlu dilakukan pra-pemrosesan. Salah satunya menggunakan *Model Output Statistics* (MOS). MOS merupakan pemodelan observasi cuaca dengan *output* NWP berbasis regresi. Observasi yang digunakan sebagai variabel respon adalah curah hujan dengan 5 kategori yakni cerah berawan, hujan ringan, hujan sedang, hujan lebat, dan hujan lebat sekali. *Output* NWP yang digunakan ada 32 variabel. Sebelumnya setiap variabel NWP dilakukan reduksi dimensi dalam sembilan grid menggunakan *Principal Component Analysis* (PCA). Metode yang digunakan untuk mengklasifikasikan curah hujan adalah klasifikasi pohon. Hasil dari PCA disimpulkan bahwa sebagian besar komponen utama yang terbentuk dari setiap variabel NWP adalah sebanyak satu komponen. Secara keseluruhan, hasil ketepatan klasifikasi curah hujan terbesar menggunakan data *testing* terletak pada stasiun pengamatan Pondok Betung. Hasil ketepatan klasifikasi data *testing* sebelum proses SMOTE pada stasiun pengamatan Citeko, Kemayoran, dan Pondok Betung yakni 100%, 85,71% dan 71,43%. Setelah proses *Synthetic Minority Oversampling Technique* (SMOTE), ketepatan klasifikasi ketiga stasiun pengamatan cenderung turun yakni 28,57%, 85,71% dan 57,14%. Berdasarkan hasil ketepatan klasifikasi data *testing* untuk setiap

stasiun pengamatan, maka pohon klasifikasi yang layak untuk klasifikasi curah hujan adalah model klasifikasi pohon optimal yang sebelum diproses menggunakan SMOTE.

***Kata Kunci*: Curah Hujan, Klasifikasi Pohon, MOS, NWP.**

ABSTRACT

MODEL OUTPUT STATISTICS USING ALGORITHM CLASSIFICATION AND REGRESSION TREES (CART) FOR CLASSIFICATION RAINFALL BY PRE-PROCESSING PRINCIPAL COMPONENT ANALYSIS (PCA)

Name of Student	: Ulul Azmi
NRP	: 1311 100 702
Departement	: Statistics
Supervisor	: Dr. Sutikno, S.Si, M.Si

ABSTRACT

A weather condition in Indonesia was announced for a period of about 24 hours with the weather forecast on the analysis of the Badan Meteorologi, Klimatologi dan Geofisika (BMKG). Since 2004, the BMKG has attempted to do research on short-term weather forecasting using weather component data Numerical Weather Prediction (NWP). But they are often biased NWP output, so it is necessary to pre-processing using Model Output Statistics (MOS). MOS is modeling weather observations with the regression-based NWP output. Observations used as the response variable is precipitation with 5 categories namely cloudy, light rain, moderate rain, heavy rain, and heavy rains all. NWP outputs are used, there are 32 variables. NWP performed before each variable dimension reduction in nine grids using Principal Component Analysis (PCA). The method used to classify the rainfall is a classification tree. The result of PCA was concluded that most of the major components formed from each variable NWP is as much as one component. Overall, the result of the classification accuracy of the heaviest rainfall using testing data is located in Pondok Betung observation stations. The results of testing the accuracy of data classification before the process SMOTE the observation station Citeko, Kemayoran, and Pondok Betung ie 100%, 85.71% and 71.43%. After the process Synthetic Minority Oversampling Technique (SMOTE), third classification accuracy of observation stations tends to fall ie 28.57%, 85.71% and 57.14%. Based on the results of testing the accuracy of data classification for each observation station, the

classification tree eligible for classification of precipitation is optimal classification tree models before being processed using SMOTE.

Keywords: Classification Trees , MOS, NWP, Rainfall

KATA PENGANTAR

KATA PENGANTAR

Alhamdulillah 'ala kulli hal. Rasa syukur penulis panjatkan atas *rahman* dan *rahiim* Allah SWT, sehingga penulis dapat menyelesaikan laporan Tugas Akhir yang berjudul "Prediksi Curah Hujan Melalui *Model Output Statistics* Menggunakan *Classification and Regression Trees* dengan *Pre-processing Principal Component Analysis*" yang disusun untuk memenuhi salah satu syarat kelulusan Program Studi S1 Jurusan Statistika FMIPA ITS.

Tugas akhir ini tidak akan selesai tanpa bantuan dan bimbingan dari berbagai pihak. Oleh karena itu, penulis menyampaikan terima kasih kepada

1. Dr. Sutikno, S.Si, M.Si selaku dosen pembimbing, atas segala bimbingan, saran, semangat, kesabaran dan waktu yang diberikan kepada penulis hingga laporan Tugas Akhir ini dapat selesai.
2. Dr. Vita Ratnasari, S.Si, M.Si selaku dosen wali yang telah membimbing dan mengarahkan selama masa perkuliahan.
3. Dr. Suhartono, M.Sc selaku Ketua Jurusan Statistika ITS.
4. Dr. Santi Wulan Purnami S.Si, M.Si selaku Koordinator Tugas Akhir Jurusan Statistika ITS.
5. Dr. rer. pol. Heri Kuswanto, S.Si., M.Si dan Dr. Purhadi, M.Sc selaku dosen penguji yang telah memberikan kritik dan saran demi kesempurnaan tugas akhir ini.
6. Seluruh dosen atas ilmu yang telah diberikan selama penulis berada di bangku kuliah dan staf Jurusan Statistika ITS yang telah membantu penulis selama pelaksanaan tugas akhir.
7. Bapak Ibu Saya, Sutrisno dan Siti Alifah atas dukungan moril dan materiil serta do'a yang tak pernah putus dan kesabaran yang diberikan.
8. Kakak dan Adik saya, Neng Tika, Ulfa dan Lilik sebagai penyemangat ketika malas melanda.
9. Seluruh teman-teman mahasiswa Statistika ITS khususnya angkatan 2011 yang selalu memberikan doa, semangat dan dorongan hingga terselesaikannya Tugas Akhir ini.

10. Teman-teman CSS MoRA ITS khususnya angkatan 2011, SATU MASA, atas segala bentuk dukungan dan semangat yang diberikan.
11. Endang Sulistiyani yang banyak membantu dalam penyempurnaan tugas akhir ini.
12. Semua sahabat yang telah memberikan do'a, semangat dan perhatian.
13. Pihak-pihak lain yang telah membantu penulis sejak pengerjaan hingga penyusunan laporan tugas akhir yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari bahwa masih banyak kesalahan dan kekurangan dalam laporan tugas akhir ini. Oleh karena itu, penulis mengharapkan kritik dan saran dari pembaca. Semoga laporan tugas akhir ini dapat bermanfaat baik bagi penulis, pembaca, maupun pihak-pihak lain.

Surabaya, Januari 2017

Penulis

DAFTAR ISI

DAFTAR ISI

Halaman

DAFTAR TABEL

DAFTAR TABEL

Halaman

(Halaman ini sengaja dikosongkan)

DAFTAR GAMBAR

DAFTAR GAMBAR

Halaman

DAFTAR LAMPIRAN

DAFTAR LAMPIRAN

Halaman

BAB I

PENDAHULUAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Di Indonesia keadaan cuaca diumumkan untuk jangka waktu sekitar 24 jam melalui prakiraan cuaca hasil analisis Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). BMKG merupakan sebuah lembaga pemerintah yang salah satu tugasnya adalah melakukan pengamatan dan prediksi terhadap unsur cuaca, diantaranya curah hujan dan sifat hujan (Paramita, 2010). Informasi cuaca telah disampaikan ke masyarakat setiap hari untuk berbagai kepentingan, seperti transportasi, kesehatan, pertanian, pembangunan intrastruktur, pariwisata dan sebagainya.

Curah hujan merupakan air yang jatuh dipermukaan tanah datar selama periode tertentu yang diukur dengan satuan tinggi milimeter (mm) di atas permukaan horizontal. Dalam penjelasan lain, curah hujan merupakan ketinggian air hujan yang terkumpul dalam tempat yang datar, tidak menguap, tidak meresap, dan tidak mengalir (BMKG, 2011). Curah hujan dan ketersediaan air tanah merupakan dua faktor utama yang saling berkaitan dalam memenuhi kebutuhan air. Namun curah hujan yang tinggi pada daerah dengan kemampuan perembesan tanah yang buruk berpotensi mengakibatkan banjir. Oleh karena itu, akurasi informasi ramalan cuaca jangka pendek seperti kejadian hujan dapat menjadi antisipasi dini terhadap dampak buruk yang diakibatkan oleh perubahan cuaca.

Dalam melakukan prediksi cuaca, saat ini BMKG berupaya menggunakan pemodelan *Numerical Weather Prediction* (NWP) yang diharapkan dapat memberikan informasi keadaan cuaca dengan akurasi yang optimal (BMKG, 2005). Model NWP merupakan sekumpulan kode komputer yang merepresentasikan persamaan atmosfer secara numerik untuk memprediksi kondisi atmosfer yang akan datang. NWP diukur dalam kombinasi lintang bujur (*grid*) tertentu sehingga menghasilkan informasi cuaca yang homogen pada beberapa daerah yang masuk dalam grid peng-

ukuran dan diukur pada skala global. Sehingga jika model NWP digunakan untuk memprediksi kondisi cuaca lokal, akan menghasilkan prediksi cuaca yang bias (Wilks, 2006). Selain itu, *output* NWP menjadi bias karena keadaan atmosfer yang tidak pasti dan terbatasnya penghitungan matematik untuk memodelkan keadaan fisik dan dinamik atmosfer (Idowu & Rautanbach, 2009). Oleh karena itu perlu dilakukan *pre-processing* data NWP sebelum digunakan untuk prediksi cuaca dengan metode statistika untuk memperbaiki hasil prediksi. Salah satu metode yang sering digunakan *Model Output Statistics* (MOS). MOS merupakan model berbasis regresi yang menghubungkan antara hasil observasi cuaca sebagai variabel respon dan *output* NWP sebagai variabel prediktor (Nichols, 2008). Pemodelan MOS memanfaat-kan data observasi cuaca dan *output* NWP.

Data NWP diambil dalam 9 grid pengukuran untuk masing-masing variabel pada setiap lokasi, sehingga memungkinkan terjadi multikolineritas karena banyaknya variabel prediktor. Guna mengatasi masalah multikolineritas pada variabel tersebut, perlu dilakukan reduksi dimensi khususnya grid variabel. Beberapa metode reduksi dimensi yang telah digunakan untuk beberapa kasus adalah *Principal Component Analysis* (PCA), *Independent Component Analysis* (ICA) (Anuravega, 2012), dan Transformasi Wavelet Diskrit (Idayati, 2014). Penelitian yang dilakukan membandingkan metode reduksi dimensi menggunakan PCA dan ICA. Hasil perbandingan PCA dan ICA menyimpulkan bahwa secara keseluruhan MOS ICA menghasilkan presisi rendah dan akurasi tinggi, sedangkan MOS PCA memiliki presisi tinggi dan akurasi rendah (Anuravega, 2012). Hasil perbandingan metode PCA dan Transformasi Wavelet Diskrit (TWD) memberi kesimpulan bahwa metode PCA menghasilkan RMSEP lebih kecil daripada metode TWD. Selain itu metode PCA mampu mengoreksi bias NWP lebih besar dibandingkan metode TWD (Idayati, 2014). Oleh karena itu reduksi dimensi pada penelitian ini menggunakan metode PCA.

Penelitian tentang prakiraan kejadian hujan menggunakan MOS pernah dilakukan oleh Prastuti pada tahun 2012 menggunakan metode regresi logistik ordinal. Hasil penelitian menyimpulkan bahwa model MOS dengan regresi logistik ordinal menghasilkan ketepatan yang cukup baik untuk klasifikasi kejadian hujan. Dalam penelitiannya, Prastuti menyatakan perlu penggunaan metode klasifikasi lainnya (Prastuti, 2013). Metode klasifikasi yang umum digunakan adalah analisis diskriminan dan regresi logistik multivariat. Namun kedua metode ini memiliki keterbatasan dalam hal pemenuhan asumsi dan kesederhanaan interpretasi. Salah satu metode yang dapat mengatasi hambatan tersebut adalah metode *Classification and Regression Trees* (CART). Metode CART merupakan metode statistika non parametrik sehingga tidak memerlukan asumsi dalam pengguna-annya (Budiyanti, 2010). Metode CART digunakan untuk meng-gambarkan hubungan antara variabel respon dan satu atau lebih variabel prediktor. Keunggulan CART dibandingkan metode klasifikasi lain adalah dapat menghasilkan tampilan grafis yang lebih mudah untuk diinterpretasikan, lebih akurat dan lebih cepat penghitungannya. Selain itu CART dapat diterapkan pada data dalam jumlah besar, variabel yang sangat banyak dan dengan skala variabel campuran melalui prosedur pemilihan biner (Statsoft, 2003).

Masalah yang sering terjadi pada klasifikasi adalah adanya *imbalance* data, dimana distribusi antara kelas mayor dan kelas minor tidak seimbang. Distribusi data yang tidak seimbang, mengakibatkan kekeliruan dalam klasifikasi kelas minor (Hairani, 2016). Salah satu metode yang mampu mengatasi masalah *imbalance* data adalah *synthetic minority oversampling technique* (SMOTE). Metode SMOTE merupakan salah satu metode *oversampling* yang bekerja dengan cara replikasi data minor (Ningrum, 2015). Sehingga diharapkan masalah *imbalance* data dapat diatasi dengan menghasilkan sampel baru dari interpolasi acak anggota minoritas yang ada (Mosley, 2013).

Pada penelitian ini, menggunakan klasifikasi pohon untuk memodelkan klasifikasi curah hujan dengan variabel NWP. Namun, terlebih dahulu dilakukan *pre-processing* reduksi dimensi grid NWP dengan metode PCA. Jika terjadi *imbalance* data, maka metode yang digunakan adalah metode SMOTE yang diharapkan dapat meningkatkan nilai akurasi hasil klasifikasi.

1.2 Rumusan Permasalahan

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah yang diangkat pada penelitian ini adalah

1. Bagaimana hasil reduksi dimensi variabel NWP dalam suatu grid pengukuran dengan metode PCA?
2. Bagaimana model klasifikasi curah hujan pada wilayah penelitian dengan metode klasifikasi pohon?
3. Bagaimana ketepatan klasfikasi curah hujan model MOS menggunakan metode klasifikasi pohon?

1.3 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut

1. Mendapatkan hasil reduksi dimensi variabel NWP dalam suatu grid pengukuran dengan metode PCA.
2. Mendapatkan model klasifikasi curah hujan dengan metode klasifikasi pohon.
3. Mengetahui ketepatan klasifikasi curah hujan di wilayah pengamatan.

1.4 Manfaat Penelitian

Manfaat yang diharapkan setelah melakukan penelitian ini adalah sebagai aplikasi ilmu statistika tentang *Model Output Statistics* (MOS) menggunakan *Classification and Regression Trees* (CART).

1.5 Batasan Masalah

Penelitian ini menggunakan data *output* NWP hasil aplikasi *Conformal Cubic Atmospheric Model* (CCAM). Data yang

digunakan adalah hasil observasi di 3 stasiun pengamatan yakni Kemayoran, Pondok Betung, dan Citeko selama 2 tahun yaitu mulai Januari tahun 2009 sampai Desember tahun 2010.

(Halaman ini sengaja dikosongkan)

BAB II

TINJAUAN PUSTAKA

BAB II
TINJAUAN PUSTAKA

2.1 *Principal Component Analysis*

Menurut Johnson (2007) konsep *Pricipal Component Analysis* (PCA) adalah pengelompokkan variabel-variabel yang berkorelasi liner menjadi 1 komponen utama, sehingga dari p variabel random (x_1, x_2, x_3,…, x_p) akan didapat k komponen utama ($k < p$) yang mewakili variabilitas variabel yang ada. Tujuan dilakukannya PCA adalah untuk mereduksi struktur hubungan variabel menjadi variabel baru dengan dimensi yang lebih kecil. Variabel baru tersebut mampu menerangkan sebagian besar varian total data dan saling bebas satu sama lain. Selanjutnya variabel baru ini dinamakan *principal component* (PC). Reduksi dimensi data pada PCA dengan cara mentransformasi variabel-variabel asli yang berkorelasi menjadi satu set variabel baru yang tidak berkorelasi, dengan tetap mempertahankan sebesar mungkin varians yang dapat dijelaskan (Johnson, 2007).

PC dapat dibentuk dari matriks kovarians maupun matriks korelasi. PC yang dibentuk dari matriks korelasi dilakukan jika variabel-variabel yang diamati mempunyai satuan pengukuran yang berbeda, maka variabel tersebut perlu distandarisasikan terlebih dahulu. Akibat adanya standarisasi data, maka matriks varians-kovarians dari data yang distandarisasi akan sama dengan matriks korelasi data sebelum distandarisasi dan besarnya total varians PC akan sama dengan banyaknya variabel asal.

Secara aljabar linier, komponen utama merupakan kombinasi linier dari p variabel acak $x_1, x_2, x_3, \dots, x_p$. Secara geometris, kombinasi linier ini merupakan sistem koordinat baru yang di dapat dari rotasi sistem semula dengan $x_1, x_2, x_3, \dots, x_p$ sebagai sumbu koordinat. Sumbu baru tersebut merupakan arah dengan variabilitas maksimum dan memberikan kovariansi yang lebih sederhana. Syarat untuk membentuk PC yang merupakan kombinasi linier dari variabel **x** agar mempunyai keragaman yang

besar adalah dengan memilih *eigenvector* $\mathbf{e_i} = (e_1, e_2, \dots, e_p)^{\mathbf{T}}$ sedemikian hingga $\mathrm{Var}(\mathbf{e_i}^{\mathbf{T}}\mathbf{x})$ maksimum dan $\mathbf{e_i}^{\mathbf{T}}\mathbf{e_i} = 1$ dan cov $(\mathbf{e_i}^{\mathbf{T}}\mathbf{x}, \mathbf{e_k}^{\mathbf{T}}\mathbf{x}) = 0$ untuk $k < i$.

PC tergantung kepada matriks varians-kovarians $\mathbf{\Sigma}$ dan matriks korelasi ρ dari $x_1, x_2, x_3, \dots, x_p$. Melalui matriks varians-kovarians diturunkan *eigenvalue* $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p$ dan *eigenvector* $\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_p$. Vektor random $\mathbf{x} = [x_1, x_2, \dots, x_p]^{\boldsymbol{T}}$ mempunyai matriks varians-kovarians $\mathbf{\Sigma}$ dengan *eigenvalue* $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0$ maka kombinasi linier utama adalah

$$\begin{aligned} PC_1 &= \mathbf{e_1}^{\mathbf{T}}\mathbf{x} = e_{11}x_1 + e_{21}x_2 + \cdots + e_{p1}x_p \\ PC_2 &= \mathbf{e_2}^{\boldsymbol{T}}\mathbf{x} = e_{12}x_1 + e_{22}x_2 + \cdots + e_{p2}x_p \\ &\vdots \\ PC_p &= \mathbf{e_p}^{\mathbf{T}}\mathbf{x} = e_{1p}x_1 + e_{2p}x_2 + \cdots + e_{pp}x_p \end{aligned} \tag{2.1}$$

dengan:

PC_1 : PC pertama, yang mempunyai varians terbesar pertama
PC_2 : PC kedua, yang mempunya varians terbesar kedua
PC_p : PC ke-*p*, yang mempunyai varians terbesar ke-*p*
x_1 : Variabel asal pertama
x_2 : Variabel asal kedua
x_p : Variabel asal ke-*p*
$\mathbf{e_p}$: *Eigenvector* variabel ke-*p*

Model *Principal Component* ke-*i* secara umum ditulis dengan:

$$PC_i = \mathbf{e_i}^{\mathbf{T}}\mathbf{x}, \text{ dimana } i = 1,2,\dots,p \tag{2.2}$$

Sehingga,

$$\mathrm{Var}(\mathrm{PC}_i) = \mathbf{e_i}^{\mathbf{T}}\mathbf{\Sigma}\mathbf{e_i} = \lambda_i \text{ dimana } i = 1,2,\dots,p \tag{2.3}$$

$$\mathrm{Cov}(\mathrm{PC}_i, \mathrm{PC}_k) = \mathbf{e_i}^{\mathbf{T}}\mathbf{\Sigma}\mathbf{e_k} = 0 \text{ untuk } i \neq k \tag{2.4}$$

Principal Component tidak berkorelasi dan mempunyai varians yang sama dengan *eigenvalue* dari $\mathbf{\Sigma}$, sehingga:

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{var}\,(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \mathrm{var}\,(PC_i) \tag{2.5}$$

Jadi persentase varians total yang dapat diterangkan oleh *Principal Component* ke-*i* adalah sebagai berikut:

$$\text{Proporsi varians ke} - i = \frac{\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_p} \qquad (2.6)$$

Apabila *Principal Component* yang diambil sebanyak k dimana $k < p$, maka:

$$\text{Proporsi varians } k \text{ PC} = \frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p} \qquad (2.7)$$

Bila PCA linier, maka menggunakan matriks kovarians dari data yang terstandarisasi karena diagonal utama matriks berisi nilai 1. Sehingga total varians populasi untuk variabel terstandarisasi adalah p. Dimana p merupakan jumlah elemen diagonal matrisk korelasi (ρ). Sehingga:

$$\text{Proporsi variansi ke} - i = \frac{\lambda_i}{p} \times 100\% \qquad (2.8)$$

Menurut Johnson dan Wichern (2007) terdapat beberapa hal yang dapat dipakai sebagai acuan dalam menentukan banyaknya PC, antara lain:

1. Melihat *scree plot*. *Scree plot* menggambarkan besarnya *eigenvalue* $\hat{\lambda}_i$. Dalam menentukan jumlah PC yang sesuai, maka bisa dilihat pada garis yang terbentuk, jika garis yang terbentuk mengalami *range* yang cukup besar maka PC sejumlah garis tersebut.
2. Apabila PC diperoleh dari matriks korelasi, maka banyaknya PC dipilih sesuai dengan banyaknya *eigenvalue* yang lebih besar dari satu.
3. Sebaiknya jumlah PC yang dipilih adalah yang mampu memberikan kumulatif persen varians 80% - 90%.

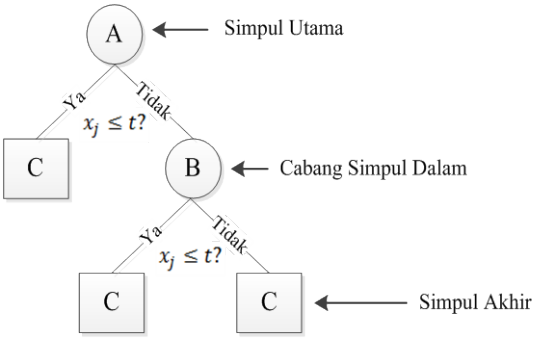
2.2 *Classification and Regression Trees* (CART)

CART merupakan salah satu metode dari teknik eksplorasi data yakni teknik pohon keputusan. Metode ini dikembangkan oleh Leo Breiman sekitar tahun 1980-an untuk melakukan analisis klasifikasi pada variabel respon nominal, ordinal maupun kontinu. Tujuan utama CART adalah mendapatkan kelompok data yang akurat sebagai penciri dari suatu pengkalsifikasian.

CART menghasilkan suatu pohon klasifikasi jika variabel responnya kategorik dan menghasilkan pohon regresi jika variabel responnya kontinu. Menurut Breiman *et al.* (1993) dalam Yusri (2008), CART merupakan metodologi statistik nonparametric dan nonlinier. Hal ini dikarenakan hasil dari CART merupakan suatu kondisi logis *if-then* dalam bentuk pohon. Sehingga tidak ada asumsi implisit bahwa hubungan antara variabel respon dan variabel prediktornya linier.

Keunggulan metode CART diantaranya dapat menghasilkan tampilan grafis yang lebih mudah untuk diinterpretasikan. Model yang dihasilkan cukup sederhana untuk dapat menjelaskan suatu amatan dikelompokkan atau diduga dalam kelompok tertentu (Statsoft, 2003). Menurut Breiman *et al.* (1993) dalam Yusri (2008), keunggulan lain dari CART adalah tidak perlu asumsi distribusi oleh semua variabel, serta algoritma yang dapat menangani data *missing* secara langsung.

Pohon klasifikasi merupakan metode penyekatan data secara berulang dan biner (*binary recursive partitioning*), karena selalu membagi kumpulan data menjadi 2 sekatan. Setiap sekatan data dinyatakan sebagai *node* (*node*). Pemilahan dilakukan pada tiap *node* samapi didapatkan suatu *node* terminal/akhir. Variabel yang memilah pada *node* utama adalah variabel terpenting dalam menduga kelas dari amatan. Lewis (2000) dalam Yusri (2008) menyebut *node* utama (*root node*) sebagai *node* induk (*parent node*), sedangkan pecahan *node* induk disebut *node* dalam (*internal nodes*). *Node* akhir yang juga disebut sebagai *node* terminal dimana sudah tidak terjadi pemilahan. Kedalaman pohon (*depth*) dihitung mulai dari *node* utama (A) berada pada kedalaman 1, sedangkan *B* berada pada kedalaman 2, begitu seterusnya sampai *node* akhir.



Gambar 2.1 Struktur Pohon Klasifikasi

Secara umum, penerapan metode CART terdiri atas 3 tahap, yakni: pembentukkan pohon klasifikasi, pemangkasan pohon klasifikasi, dan penentuan pohon klasifikasi optimal.

2.2.1 Pembentukkan Pohon Klasifikasi

Proses pembentukkan pohon klasifikasi terdiri atas 3 tahap yakni pemilihan pemilah, penentuan *node* terminal, dan penandaan label kelas. Proses pembentukan pohon klasifikasi dibutuhkan data *learning* sehingga perlu dicari terlebih dahulu metode terbaik untuk pembetukan pohon terbaik dengan ketepatan klasifikasi tertinggi pada data *testing*. Data dibagi menjadi data *learning* (L_1) dan data *testing* (L_2).

1) Pemilihan Pemilah (*Classifier*)

Pada tahap ini sampel data *training* yang masih bersifat heterogen digunakan untuk pembentukan pohon klasifikasi. Kemudian dicari pemilah dari setiap variabel dalam *node* yang menghasilkan penurunan tingkat keheterogenan paling tinggi. Tingkat keheterogenan diukur berdasarkan nilai *impurity*. *Impurity measure i(t)* merupakan pengukuran tingkat keheterogenan suatu kelas dari suatu *node* tertentu dalam pohon klasifikasi yang dapat membantu menemukan fungsi pemilah yang optimal. Beberapa fungsi *impurity* yang dapat digunakan adalah Indeks Gini, Indeks Informasi, Indeks *Twoing* dan Indeks Entropi. Fungsi *impurity* yang umum digunakan adalah Indeks Gini karena

proses perhitungan yang sederhana dan sesuai diterapkan dalam berbagai kasus. Ide dasar dari Indeks Gini adalah memisahkan kelas dengan anggota paling besar atau kelas terpenting dalam *node* tersebut terlebih dahulu. Pemilah terbaik dipilih dari semua kemungkinan pemilahan pada setiap variabel prediktor berdasarkan pada nilai penurunan keheterogenan tertinggi (Breiman, 1993). Fungsi *impurity* Indeks Gini dituliskan dalam

$$i(t) = \sum_{i \neq j} p(i|t)\, p(j|t) \tag{2.9}$$

$p(i|t)$: Proporsi kelas *i* pada *node t*.
$p(j|t)$: Proporsi kelas *j* pada *node t*.

Karena beberapa kelebihan Indeks Gini, maka fungsi *impurity* yang digunakan pada penelitian ini adalah fungsi *impurity* Indeks Gini.

2) Penentuan *Node* Terminal

Suatu *node t* akan menjadi *node* terminal atau tidak dilihat dari kondisi *node* yang memenuhi salah satu kriteria berikut (Breiman, 1993).

b. Hanya ada satu pengamatan (*n*=1) dalam tiap *node* anak atau adanya batasan minimum *n* pengamatan yang diinginkan peneliti.
c. Semua pengamatan dalam setiap *node* anak mempunyai distribusi yang identik terhadap variabel prediktor sehingga tidak mungkin untuk dipilih lagi.
d. Adanya batasan jumlah level atau tingkat kedalaman pohon maksimal yang ditetapkan peneliti.

Apabila struktur pohon telah terbentuk mulai dari *node* utama sampai dengan *node* terminal dimana sudah tidak lagi ditemukan *node* yang perlu dipilah lagi maka pohon klasifikasi maksimal telah terbentuk. Pohon klasifikasi maksimal merupakan pohon klasifikasi yang memiliki jumlah *node* paling banyak (Breiman, 1993).

3) Penandaan Label Kelas

Penandaan label kelas pada *node* terminal dilakukan berdasarkan aturan jumlah terbanyak. Label kelas *node* terminal *t*

adalah j_0 yang memberi nilai dugaan kesalahan pengklasifikasian *node t* terbesar. Proses pembentukan pohon klasifikasi berhenti saat hanya terdapat satu pengamatan dalam tiap *node* anak atau adanya batasan minimum *n*. Semua pengamatan dalam tiap *node* anak adalah identik dan adanya batasan jumlah kedalaman pohon maksimal (Breiman, 1993).

$$p(j_0|t) = max_j p(j|t) = max_j \frac{N_j(t)}{N(t)} \qquad (2.10)$$

dimana

$p(j|t)$: Proporsi kelas *j* pada *node*

$N_j(t)$: Jumlah pengamatan kelas *j* pada *node t*

$N(t)$: Jumlah pengamatan pada *node t*

2.2.2 Pemangkasan Pohon Klasifikasi (*Prunning*)

Pemangkasan dilakukan pada bagian pohon yang kurang penting, sehingga akan didapatkan pohon klasifikasi yang optimal. Pemangkasan didasarkan pada suatu penilaian ukuran sebuah pohon tanpa mengorbankan kebaikan ketepatan melalui pengurangan *node* pohon sehingga dicapai ukuran pohon yang layak. Ukuran pemangkasan yang digunakan untuk memperoleh ukuran pohon yang layak disebut *Cost complexity minimum* (Lewis 2000 dalam Yusri (2008)).

Jika T diperoleh dari T_{max} sebagai hasil dari pemangkasan suatu *branch*, maka T disebut *pruned subtree* dari T_{max} yang dinotasikan dengan $T < T_{max}$. dimana $T < T_{max}$ memiliki *root node* yang sama. Metode yang digunakan untuk pemangkasan pohon berdasarkan pada minimal *cost complexity pruning*.

$$R(T) = \sum_{t \in T} R(t) \qquad (2.11)$$

$R(T)$ merupakan *tree resubstitution cost*, sedangkan $R(t)$disebut *node misclassification cost.*

Proses pemangkasan pohon dimaksudkan untuk mengatasi *overfitting* dan penyederhanaan interpretasi. Pemangkasan dilakukan dengan memotong pohon maksimal (T_{max}) menjadi beberapa pohon klasifikasi (T) yang ukurannya lebih kecil (*subtrees*).

Diketahui *subtree* $T < T_{max}$ didefinisikan komplesitas dari *subtree* ini adalah $|\tilde{T}|$, yakni banyaknya terminal *node* yang dimiliki pohon *T*. Nilai $\alpha \geq 0$ merupakan *complexity parameter* dan $R_\alpha(T)$ merupakan *cost complexity measure*, maka:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{2.12}$$

dimana

$R(T)$: *Tree resubstitution cost* (Proporsi kesalahan pada sub pohon)

α : *Complexity parameter*

$|\tilde{T}|$: Ukuran banyaknya *node* terminal pohon *T*

Secara umum tahapan pada proses pemangkasan pohon adalah sebagai berikut.

1. Membentuk pohon klasifikasi maksimal T_{max} kemudian diambil *node* anak kanan t_R dan *node* anak kiri t_L dari T_{max} yang dihasilkan dari pemilahan *node* induk *t*.
2. Jika diperoleh dua *node* anak dan *node* induknya yang memenuhi persamaan $R(t) = R(t_L) + R(t_R)$, maka *node* anak t_L dan $t_R$dipangkas. Hasilnya merupakan pohon T_1 yang memenuhi kriteria $R(T_1) = R(T_{max})$.
3. Ulangi langkah 2 sampai tidak ada lagi pemangkasan yang mungkin. Hasil proses pemangkasan adalah suatu barisan menurun dan tersarang dari pohon bagian yaitu $T_1 > T_2 > \cdots > \{t_1\}$ dengan $T_1 < T_{max}$ dan suatu barisan menaik dari parameter *cost complexity*, yaitu $\alpha_1 = 0 < \alpha_2 < \alpha_3 < \cdots$

2.2.3 Penentuan Pohon Klasifikasi Optimal

Ukuran pohon yang besar akan mengakibatkan nilai komplesitas yang tinggi karena struktur data yang digambarkan cenderung kompleks. Sehingga perlu dipilih pohon optimal yang berukuran proporsional tetapi memberikan nilai penduga pengganti cukup kecil. Terdapat 2 jenis penduga pengganti yakni, penduga sampel uji (*test sample estimate*) dan penduga *cross validation V-fold.* Penelitian ini menggunakan penduga *cross*

validation V-fold karena data penelitian yang digunakan kurang dari 3000.

1) Penduga *Cross Validation V-Fold*

Penduga ini sering dilakukan apabila pengamatan yang ada tidak cukup besar. *Cross validation* membagi data secara acak menjadi *V* subset yang berukuran relatif sama. Salah satu subset dicadangkan sebagai data *testing* dan subset-subset sisanya digabung dijadikan sebagai data *learning* dalam prosedur pembentukan model. Seluruh prosedur pembentukan model diulang *V* kali, dengan subset berbeda dari data setiap kali melakukan pembentukan pohon (Lewis, 2000). Nilai *V* yang sering dipakai dan dijadikan standar adalah 10. Karena hasil dari berbagai percobaan ekstensif dan pembuktian teoritis, menunjukkan bahwa *cros validation 10-fold* adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat.

Cross validation v-fold estimation untuk T_k yang menggunakan pengamatan L dalam membentuk deretan pohon $\{T_k\}$ adalah sebagai berikut.

$$R^{cv}\big(T_k(\alpha)\big) = \frac{1}{N}\sum_{i,j} C(i|j)\, N_{ij} \qquad (2.13)$$

dimana

$R^{cv}\big(T_k(\alpha)\big)$: Total proporsi *t cross validation v-fold estimation*

$C(i|j)$: Jumlah proporsi ke-*i* dan ke-*j* dari keseluruhan data pengamatan

N_{ij} : Jumlah kelas ke-*i* dan ke-*j* dari keseluruhan data pengamatan

Pohon klasifikasi optimal yang dipilih yaitu T_k dengan $R^{cv}(T_k) = \min_k R^{cv}(T_k)$.

2.3 Ukuran Ketepatan Klasifikasi

Salah satu cara yang dapat digunakan untuk mengukur ketepatan klasifikasi diantaranya melalui perhitungan *Apparent Error Rate* (APER) dan *total accuracy rate* (1-APER*)*. Menurut Johnson dan Wichern (2007) *Apparent Error Rate* (APER) merupakan proporsi observasi yang diprediksi secara tidak benar

(ukuran kesalahan klasifikasi total). *Total accuracy rate* (1-APER) merupakan proporsi observasi yang diprediksi secara benar (ukuran ketepatan klasifikasi total). *Crosstab* untuk menghitung ketepatan klasifikasi ditunjukkan dalam Tabel 2.1 berikut.

Tabel 2.1 *Crosstab* Ketepatan Klasifikasi

Kelompok Aktual Variabel Y	**Kelompok Prediksi Variabel Y**					**Jumlah Observasi**
	1	**2**	**3**	**4**	**5**	
1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_1
2	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	n_2
3	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	n_3
4	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	n_4
5	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}	n_5

dengan,

n_{11} : Frekuensi variabel Y pada kategori 1 yang tepat diprediksikan sebagai variabel Y Kategori 1

n_{21} : Frekuensi variabel Y pada kategori 2 yang tepat diprediksikan sebagai variabel Y Kategori 1

n_{12} : Frekuensi variabel Y pada kategori 1 yang tepat diprediksikan sebagai variabel Y Kategori 2

n_{22} : Frekuensi variabel Y pada kategori 2 yang tepat diprediksikan sebagai variabel Y Kategori 2

n_{55} : Frekuensi variabel Y pada kategori 5 yang tepat diprediksikan sebagai variabel Y Kategori 5

n_1 : Frekuensi variabel Y pada kategori 1

n_2 : Frekuensi variabel Y pada kategori 2

n_3 : Frekuensi variabel Y pada kategori 3

n_4 : Frekuensi variabel Y pada kategori 4

n_5 : Frekuensi variabel Y pada kategori 5

Berikut perhitungan untuk APER, dan *total accuracy rate* (1-APER).

$$\text{APER} = \frac{n_{12}+n_{13}+n_{14}+n_{15}+\cdots+n_{51}+n_{52}+n_{53}+n_{54}+n_{55}}{n_1+n_2+n_3+n_4+n_5} \qquad (2.14)$$

Total Accuracy Rate = 1 - APER (2.15)

2.4 *Synthetic Minority Oversampling Technique* (SMOTE)

Algoritma *Synthetic Minority Oversampling Technique* (SMOTE) pertama kali ditemukan oleh Chawla (2002). SMOTE merupakan salah satu metode *oversampling*, yaitu metode pengambilan sampel untuk meningkatkan jumlah data pada kelas minor dengan cara mereplikasi jumlah data pada kelas minor secara acak. Pendekatan ini bekerja dengan membuat *synthetic* data, yakni data replikasi dari data minor. Mirip dengan metode *clustering*, teknik ini sangat sederhana dan mudah untuk diimplementasikan. Metode SMOTE bekerja dengan mencari *k-nearest neighbor* (tetangga terdekat) untuk *oversampling* kelas minoritas. Tetangga terdekat dipilih berdasarkan jarak *euclidean* antara kedua data. Diharapkan masalah *overfitting* dapat diatasi dengan menghasilkan *instances* baru dari interpolasi acak anggota minoritas yang ada (Mosley, 2013).

Tetangga terdekat dipilih berdasarkan jarak *euclidean* antara data. Misalkan diberikan dua data dengan *p* dimensi yaitu $\boldsymbol{x}^{\boldsymbol{T}} = [x_1, x_2, \dots, x_p]$ dan $\boldsymbol{y}^{\boldsymbol{T}} = [y_1, y_2, \dots, y_p]$ maka jarak *euclidean* $d(x, y)$ antara kedua vektor data adalah sebagai berikut,

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \quad (2.16)$$

Sedangkan *synthetic* data dilakukan dengan menggunakan persamaan berikut.

$$\left(x_{syn}\right) = x_i + (x_{knn} - x_i) \times \beta;\ i = 1,2, \dots, n \quad (2.17)$$

dengan,

x_{syn} : Data hasil replikasi

x_i : Data yang akan direplikasi

x_{knn} : Data yang memiliki jarak terdekat dari data yang akan direplikasi

β : Bilangan random antara 0 sampai 1

Tahapan yang perlu dilakukan pada algoritma SMOTE adalah sebagai berikut

1. Mencari tetangga terdekat (x_{knn}) untuk setiap data pada kelas minor yang akan direplikasi menggunakan jarak *euclidean.* Kemudian dipilih jarak terpendek dari hasil perhitungan jarak *euclidean.*
2. Menghitung *synthetic data* (x_{syn}) menggunakan persamaan 2.17

2.5 *Numerical Weather Prediction* (NWP)

NWP diukur dalam domain lokasi atau grid yang tinggi, yaitu antara 7 sampai 60 km, dengan skala sebesar itu NWP akan memberikan informasi cuaca yang homogen pada daerah grid tersebut. Kondisi cuaca skala kecil atau skala lokal kurang tereprentasikan dengan baik. Oleh karenanya, *output* NWP memiliki sifat bias dalam meramalkan kondisi cuaca lokal karena diukur dengan domain yang tinggi. Selain itu *output* NWP juga bersifat deterministik dan tidak bisa secara penuh menjelaskan proses stokastik cuaca. Sehingga perlu dilakukan pemrosesan secara statistik (*statistical post-processing*) agar mampu menjelaskan ketidakpastian tersebut (Wilks, 2006).

Conformal Cubic Atmospheric Model (CCAM) adalah salah satu model aplikasi yang menghasilkan produk NWP. CCAM pertama kali dikembangkan oleh CSIRO (*Commonwealth Scientific and Industrial Research Organization*) Australia yang sebelumnya menggunakan *Division of Atmospheric Research Limited Area Model* (DARLAM). Kemudian CCAM diterapkan di Indonesia pada tahun 2007. *Input* yang diperlukan oleh CCAM adalah AVN/GFS. AVN/GFS adalah model *spectral* untuk prediksi cuaca global yang dijalankan oleh *National Centers for Enviromental Prediction* (NCEP). Model ini dapat memprediksi keadaan cuaca seluruh dunia sampai 2 minggu ke depan (BMG, 2008).

Menurut Raible *et al*. (1998) dalam Arifianto 2008, secara umum model-model NWP cukup baik dalam peramalan jangka pendek (*short-term forecasting*) sampai dengan 24 jam kedepan.

NWP dicatat pada grid (kombinasi lintang-bujur) tertentu dengan deskripsi sebagai berikut.

1. Variabel NWP diantaranya *Surface Pressure tendency* (dpsdt), *Water Mixing Ratio* (mixr), *Geopotential Height* (Z), *Temperature* (T), *Relative Humidity* (Rh), komponen U-V (komponen angin timur dan barat), *Mean Sea Level Pressure* (psl), *Vertical Velocity* (omega), *Maximum Screen Temperature* (tmaxscr), *Minimum Screen Temperature* (tminscr).
2. Level tekanan: 1000 mb, 950 mb, 925 mb, 900 mb, 850 mb, 800 mb, 700 mb, 600 mb, 500 mb, 400 mb, 350 mb, 300 mb dan 200mb. Level ketinggian: permukaan, 2 meter, dan 10 meter. NWP diukur pada level tekanan tertentu, dan dapat diukur pada level ketinggian: permukaan laut, 2 meter, dan 10 meter di atas permukaan laut.
3. Ramalan NWP dilakukan setiap 6 jam sekali, yaitu pada jam ke-00, 06, 12, 18, 24, 36, 42, 48, 54, 60, 66, dan 72.
4. Resolusi: grid lintang bujur 1.5^o x 1.5^o. NWP diukur pada grid poin yang luas dengan ukuran lintang bujur tertentu.

Hasil dari prakiraan NWP dengan resolusi tinggi di suatu tempat (grid) seringkali menghasilkan bias yang besar terutama untuk wilayah dengan topografi dan tutupan vegetasi yang kompleks.

2.6 *Model Output Statistics (MOS)*

Hasil peramalan cuaca dengan menggunakan model *Numerical Weather Prediction* (NWP) pada suatu lokasi tertentu dengan resolusi tinggi seringkali bias. NWP yang diukur secara global pada lokasi dengan domain yang tinggi sulit untuk meramalkan keadan cuaca lokal sehingga hasil ramalan cuaca yang dihasilkan adalah bias. Selain itu model NWP menghasilkan ramalan cuaca yang bias dikarenakan keadaan atmosfir yang tidak pasti dan terbatasnya perhitungan matematik untuk memodelkan keadaan fisik dan dinamika atmosfir. Oleh karena itu diperlukan suatu pemrosesan secara statistik (*statistical post processing*)

yang berguna untuk meningkatkan keakuratan hasil ramalan cuaca menggunakan model NWP. Salah satu metode yang dapat digunakan adalah *Model Output Statistics* (MOS), metode ini menentukan hubungan statistik antara prediktan dan variabel dari model numerik pada beberapa proyeksi waktu (Idowu & Rautanbach, 2009).

MOS pertama kali diperkenalkan dan dikembangkan oleh Glahn dan Lowry pada tahun 1969 dan dipublikasikan pada tahun 1972. MOS merupakan model berbasis regresi yang menghubungkan antara variabel respon *y* hasil observasi cuaca, dengan variabel prediktor *x* parameter NWP (Nichols, 2008). Metode regresi yang digunakan dapat menggunakan pendekatan parametrik ataupun nonparametrik tergantung dari struktur dan pola data.

Menurut Wilk (2006) Secara umum persamaan matematis MOS adalah sebagai berikut.

$$\hat{y}_t = \hat{\hat{f}}_{MOS}(\mathbf{x}_t) \tag{2.18}$$

dimana:

$\hat{y}_t$: Ramalan cuaca saat *t*

$\mathbf{x}_t$: Variabel parameter NWP pada waktu *t*

MOS akan menghasilkan ramalan yang optimal jika memenuhi syarat berikut:

1. Periode data untuk *training* (verifikasi) model seharusnya sepanjang mungkin (beberapa tahun). Data *training* yang dimaksud adalah data yang digunakan dalam pembangunan model regresi.
2. Model yang terbentuk seharusnya tidak berubah pada kondisi ekstrim selama verifikasi model.
3. Pada tahap validasi model, MOS seharusnya dapat diaplikasikan dan tidak berubah modelnya. Validasi model dimaksudkan untuk menguji keandalan model yang sudah dibangun dengan menggunakan data independen. Salah satu cara menvalidasi adalah validasi silang (*cross validation* MOS), yaitu mempartisi data (misal setiap bagian 10%)

kemudian model regresi dibentuk dengan data 90% (untuk verifikasi) dan sisanya digunakan untuk validasi. Proses ini dilakukan secara berulang sebanyak 10 kali dengan sekumpulan data yang berbeda (BMKG, 2006).

Menurut Maini dan Kumar (2004) dalam Priambudi (2006), kombinasi linier terbaik antara variabel respon dan variabel prediktor (data NWP) terletak pada 9 grid di sekitar stasiun pengamatan. Model MOS memiliki kemampuan untuk melakukan peramalan hingga 72 jam kedepan.

2.7 Konsep Dasar Curah Hujan

Dalam ilmu meteorologi, hasil dari kondensasi uap air di atmosfer disebut sebagai presipitasi yang terjadi ketika atmosfer menjadi jenuh dan air terkondensasi. Presipitasi yang mencapai permukaan bumi salah satunya adalah dalam bentuk hujan. Curah hujan mempunyai variabilitas yang besar dalam ruang dan waktu yang mengakibatkan adanya fluktuasi curah hujan. Hujan merupakan gejala atau fenomena cuaca yang dipandang sebagai variabel tak bebas karena terbentuk dari proses berbagai unsur. Curah hujan adalah air yang jatuh dipermukaan tanah datar selama periode tertentu yang diukur dengan satuan tinggi milimeter (mm) di atas permukaan horizontal. Dalam penjelasan lain, curah hujan merupakan ketinggian air hujan yang terkumpul dalam tempat yang datar, tidak menguap, tidak meresap, dan tidak mengalir. Curah hujan 1 milimeter, artinya dalam luasan $1m^2$ pada tempat yang datar tertampung air setinggi satu milimeter atau tertampung air sebanyak satu liter. Jumlah curah hujan dalam satu dasarian (rentang waktu selama 10 hari) lebih dari 50 milimeter dan diikuti oleh beberapa dasarian berikutnya ditetapkan sebagai permulaan musim hujan (BMKG, 2011). Berdasarkan intensitasnya, curah hujan diklasifikasikan menjadi lima seperti pada Tabel 2.2 (Sumber: BMKG, 2006).

Tabel 2.2 Klasifikasi Intensitas Curah Hujan

Klasifikasi Hujan	Intesitas Curah Hujan (mm/hari)
Cerah berawan	Curah Hujan ≤ 0,1
Hujan ringan	0,1 < Curah Hujan ≤ 20
Hujan sedang	20 < Curah Hujan ≤ 50
Hujan lebat	50 < Curah Hujan ≤ 100
Hujan lebat sekali	Curah Hujan > 100

Alat yang digunakan untuk mengukur curah hujan berbentuk silinder yang biasa diletakkan di tempat yang terbuka dan tidak tertutup oleh pohon dan gedung. Pencatatan dilakukan setiap hari, biasanya pukul 09.00 dan hasil pencatatan dicatat sebagai curah hujan hari terdahulu (Idayati, 2014).

2.8 Penelitian Sebelumnya

Beberapa penelitian dengan menggunakan metode MOS untuk meramal cuaca jangka pendek dengan menggunakan berbagai pendekatan regresi, seperti regresi linier berganda (Idowu, 2008), *Projection Pursuit Regression* (Safitri, 2012), regresi logistik ordinal (Prastuti, 2013) dan SIMPLS (Septiana, 2014). Penelitian terkait pembandingan metode reduksi pernah dilakukan oleh Anuravega (2012) dan Idayati (2014).

Dari hasil penelitian Idowu (2008) mennyimpulkan bahwa model MOS dapat memperbaiki hasil ramalan NWP sebesar 76% dengan variabel respon suhu dan kelembapan. Hasil penelitian Safitri (2012) mendapatkan nilai RMSEP model MOS secara konsisten lebih kecil daripada model NWP untuk semua variabel respon yang digunakan di 4 stasiun pengamatan. Penelitian berkaitan kejadian hujan menggunakan MOS pernah dilakukan oleh Prastuti pada tahun 2013. Penelitian ini menggunakan regresi logistik ordinal, karena respon dikategorikan menjadi 5 yaitu: cerah berawan, hujan ringan, hujan sedang, hujan lebat, dan hujan lebat sekali sebagaimana kategori yang dilakukan oleh BMKG. Data yang digunakan yakni data NWP harian dengan periode 01 Januari 2009 sampai 31 Desember 2010. Model dibangun berdasarkan persamaan regresi logistik ordinal.

Ketepatan klasifikasi kejadian hujan terbesar terdapat pada stasiun pengamatan Tangerang. Sedangkan hasil ketepatan klasifikasi terkecil untuk data *training* dan *testing* terdapat pada stasiun yang berbeda, yaitu Darmaga dan Curug. Model MOS dengan regresi logistik ordinal menghasilkan ketepatan yang cukup baik untuk klasifikasi kejadian hujan.

Septiana (2014) melakukan penelitian model MOS menggunakan metode regresi *Statistically Inspired Modification of Partial Least Square* (SIMPLS). Observasi cuaca yang digunakan sebagai variabel respon adalah T_{MAX}, T_{MIN}, dan RH, sedangkan parameter NWP yang digunakan sebanyak 18 variabel. Sebagian besar komponen utama yang terbentuk dari setiap variabel NWP adalah sebanyak satu komponen. Hasil penelitian menyimpulkan bahwa validasi model SIMPLS dengan kriteria RMSEP menunjukkan bahwa RMSEP untuk T_{MAKS} di empat stasiun berkriteria sedang. Nilai %IM untuk prediksi T_{MIN} mencapai 89,75%, yang artinya model SIMPLS dapat meng-koreksi bias NWP sebesar 89,75%.

Dalam penelitiannya, Anuravega tahun 2012 membandingkan metode reduksi dimensi menggunakan metode *Principal Component Analysis* (PCA) dan membandingkan metode reduksi dimensi menggunakan metode *Indpendent Component Analysis* (ICA). Hasil penelitian tersebut menyimpulkan bahwa secara keseluruhan MOS ICA menghasilkan presisi rendah dan akurasi tinggi, sedangkan MOS PCA memiliki presisi tinggi dan akurasi rendah. Penelitian serupa juga pernah dilakukan oleh Idayati (2014) menggunakan metode *Principal Component Analysis* (PCA) dan Tranformasi Wavelet Diskrit (TWD). Hasil penelitian menyimpulkan bahwa metode PCA menghasilkan RMSEP lebih kecil dibandingkan metode TWD. Selain itu metode PCA mampu mengoreksi bias NWP lebih besar dibandingkan metode TWD.

Penelitian terkait klasifikasi pernah dilakukan (Yusri, 2008) menggunakan metode CART. Penelitian ini bertujuan melihat variabel yang dapat mempengaruhi status daerah kabupaten di Indonesia berdasarkan variabel yang telah ditetapkan oleh

KNPDT (Kementerian Negara Pembangunan Daerah Tertinggal). Pada tahun 2014, Febti melakukan penelitian klasifikasi pengangguran terbuka menggunakan CART. Hasilnya dapat diketahui faktor yang mempengaruhi pengangguran terbuka adalah jenis kelamin, pendidikan terakhir, usia, status dalam rumah tangga, dan status perkawinan. Ketepatan klasifikasi yang dihasilkan sebesar 78,90 %.

BAB III
METODOLOGI PENELITIAN

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Penelitian ini menggunakan data sekunder yakni data *output* NWP *Conformal Cubic Atmospheric Model* (CCAM) periode 01 Januari 2009 sampai 31 Desember 2010, yang didapat dari NWP *Arpeg Tropic Products Meteo Franc*. Penelitian ini juga menggunakan data curah hujan harian wilayah Jabodetabek yang diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG). Terdapat 3 wilayah pengamatan yang menjadi wilayah penelitian yakni Stasiun Pengamatan Kemayoran, Pondok Betung dan Citeko dengan Lintang Bujur pada Tabel 3.1. Ketiga stasiun pengamatan tersebut dipilih karena ketiga stasiun tersebut memiliki catatan pengamatan yang cukup lengkap.

Tabel 3.1 Wilayah Stasiun Pengamatan

No	Kabupaten	Nama Stasiun	Lintang	Bujur
1	DKI Jakarta	Stasiun Kemayoran	-6.18	106.85
2	Tangerang	Stasiun Pondok Betung	-6.25	106.76
3	Bogor	Stasiun Citeko	-6.42	106.85

3.2 Variabel Penelitian

Variabel respon yang digunakan dalam penelitian adalah curah hujan harian dengan variabel prediktor berupa *output* NWP yang merupakan aplikasi model CCAM dengan parameter yang disajikan pada Tabel 3.2.

Tabel 3.2 Parameter *Output* NWP

No	Nama Variabel	Level
1.	*Surface Pressure Tendency* (dpsdt)	Permukaan
2.	*Water Mixing Ratio* (mixr)	1, 2, dan 4
3.	*Vertical Velocity* (omega)	1, 2, dan 4
4.	PBL *depth* (pblh)	Permukaan
5.	*Surface Pressure* (ps)	Permukaan
6.	*Mean Sea Level Pressure* (psl)	Permukaan
7.	*Screen Mixing Ratio* (qgscrn)	Permukaan
8.	*Relative Humidity* (rh)	1, 2, dan 4
9.	*Precipitation* (rnd)	Permukaan
10.	*Temperature*	1, 2, dan 4
11.	*Maximum Screen Temperature* (tmaxcr)	Permukaan
12.	*Minimum Screen Temperature* (tmincr)	Permukaan
13.	*Pan Temperature* (tpan)	Permukaan
14.	*Screen Temperature* (tscrn)	Permukaan
15.	*Zonal Wind* (u)	1, 2, dan 4
16.	*Friction Velocity* (ustar)	Permukaan
17.	*Meridional Wind* (v)	1, 2, dan 4
18.	*Geopotential Height* (zg)	1, 2, dan 4

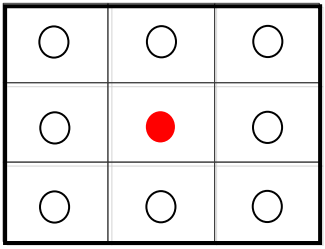
Parameter NWP yang akan digunakan pada masing-masing lokasi pengamatan ada sebanyak 18 parameter. Dimana 11 parameter diukur pada level permukaan yaitu pada ketinggian ±2m, dan 7 parameter lainnya diukur pada tiga level tekanan yang berbeda yaitu 1, 2, dan 4. Level 1 merupakan level saat tekanan 1000 mb (milibar), level 2 merupakan level saat tekanan 950 mb dan level 4 merupakan level saat tekanan 850 mb. Jadi, jumlah parameter keseluruhan yang digunakan sebanyak 32 parameter. Kemudian dari 32 parameter tersebut, masing-masing parameter akan diukur pada sembilan grid (3 x 3) pengukuran yang terdekat dari lokasi stasiun pengamatan. Masing-masing parameter akan mempunyai korelasi yang kuat dengan dirinya sendiri karena diukur pada 9 grid pengukuran. Sedangkan korelasi kuat juga terjadi antar parameter NWP karena data NWP berdimensi tinggi.

Berikut merupakan definisi dari masing-masing *output* NWP yang digunakan sebagai variabel prediktor dalam penelitian ini.

1. *Surface Pressure Tendency* (dpsdt) atau kecenderungan tekanan udara merupakan suatu indikasi dari arah dan intensifikasi suatu disturbansi siklon.
2. *Water Mixing Ratio* merupakan rasio jumlah uap air yang ada di udara.
3. *Vertical Velocity* (omega) merupakan ukuran kecepatan angin vertikal.
4. *Planetrary Boundary Layer* (pblh) adalah suatu skala ketinggian yang sangat penting dalam model atmosfer untuk mendeskripsikan vertikal mixing dari turbulansi dan konveksi awan kumulus.
5. *Surface Pressure* (ps) atau tekanan udara diukur berdasarkan tekanan gaya pada permukaan dengan luas tertentu. Alat pengukur tekanan udara disebut barometer. Tekanan udara berkurang dengan bertambahnya ketinggian.
6. *Mean Sea Level Pressure* (psl) adalah suhu rata-rata di atas permukaan laut.
7. *Relative Humidity* (rh) atau kelembapan adalah konsentrasi uap air di udara. Alat untuk mengukur kelembapan disebut higrometer.
8. *Precipitation* (rnd) dikenal sebagai salah satu kelas hydrometeors, yang merupakan fenomena air di atmosfer. rnd merupakan setiap produk dari kondensasi uap air di atmosfer yang jatuh karena gravitasi
9. *Temperature* atau Suhu udara adalah derajat panas dan dingin udara di atmosfer. Alat untuk mengukur suhu udara disebut termometer. Pengukuran suhu udara biasanya dinyatakan dalam skala Celcius (C), Reamur (R), Farenheit (F), atau Kelvin (K). Suhu udara memiliki hubungan berbanding terbalik dengan tekanan udara.
10. *Maximum Screen Temperature* (tmaxscr) merupakan suhu tertinggi yang terukur pada grid-grid.

11. *Minimum Screen Temperature* (tminscr) merupakan suhu terendah yang terukur pada grid-grid.
12. *Screen Temperature* (tscrn) adalah derajat panas dan dingin udara pada grid-grid.
13. *Zonal Wind* (u) atau komponen U adalah komponen angin yang bergerak dengan arah barat-timur.
14. *Friction Velocity* (ustar) atau kecepatan gesekan, adalah bentuk tegangan geser dan dapat ditulis dalam satuan kecepatan.
15. *Meridional Wind* (v) atau komponen V adalah komponen angin yang bergerak dengan arah utara-selatan.
16. *Geopotential Height* (zg) adalah koordinat vertikal yang direferensikan ke permukaan laut bumi atau suatu penyesuaian terhadap tinggi geometris dengan menggunakan variasi gravitasi dengan garis lintang dan ketinggian.

Terdapat 18 parameter NWP CCAM yang digunakan pada masing-masing wilayah pengamatan. Sebanyak 11 parameter diukur pada level permukaan yakni pada ketinggian ±2 meter dan 7 parameter yang diukur pada 3 level tekanan yang berbeda yaitu 1, 2, dan 4. Dimana level 1 merupakan keadaan saat tekanan 1000 mb (milibar), level 2 saat tekanan 950 mb dan level 4 saat tekanan 850 mb. Jumlah parameter keseluruhan menjadi 32 parameter, kemudian masing-masing parameter akan diukur pada 9 grid pengukuran terdekat dari lokasi stasiun pengamatan. Resolusi grid yang digunakan adalah $1{,}5^{o}$x$1{,}5^{o}$. Proyeksi pengukuran variabel NWP dalam grid 3x3 ditunjukkan pada Gambar 3.1.



Gambar 3.1 Pengukuran NWP dalam grid 3x3

Titik merah pada Gambar 3.1 menunjukkan grid terdekat pada lokasi stasiun pengamatan, sedangkan kotak yang berwarna hitam merupakan kombinasi grid di sekitar lokasi pengamatan. Oleh karena itu, masing-masing variabel akan mempunyai korelasi yang kuat dengan dirinya sendiri karena diukur pada 9 grid pengukuran. Sedangkan antar variabel prediktor juga akan mempunyai korelasi yang kuat, hal ini dikarenakan data NWP berdimensi tinggi. Karena memiliki dimensi yang cukup besar, maka perlu dilakukan reduksi dimensi pada grid pengukuran variabel NWP menggunakan *Principal Component Analysis* (PCA).

Hasil dari reduksi PCA berupa beberapa komponen utama kemudian digunakan sebagai variabel prediktor untuk membangun klasifikasi pohon. Sedangkan variabel respon curah hujan akan diklasifkasikan menjadi 5 kategori, dengan kriteria sebagai berikut:

Tabel 3.3 Klasifikasi Curah Hujan Menurut Intensitasnya

Klasifikasi Hujan	Intesitas curah hujan (mm/hari)
Cerah berawan	Curah hujan ≤ 0,1
Hujan ringan	0,1 < Curah hujan ≤ 20
Hujan sedang	20< Curah hujan ≤ 50
Hujan lebat	50 < Curah hujan ≤ 100
Hujan lebat sekali	Curah hujan > 100

(Sumber: BMKG, 2006)

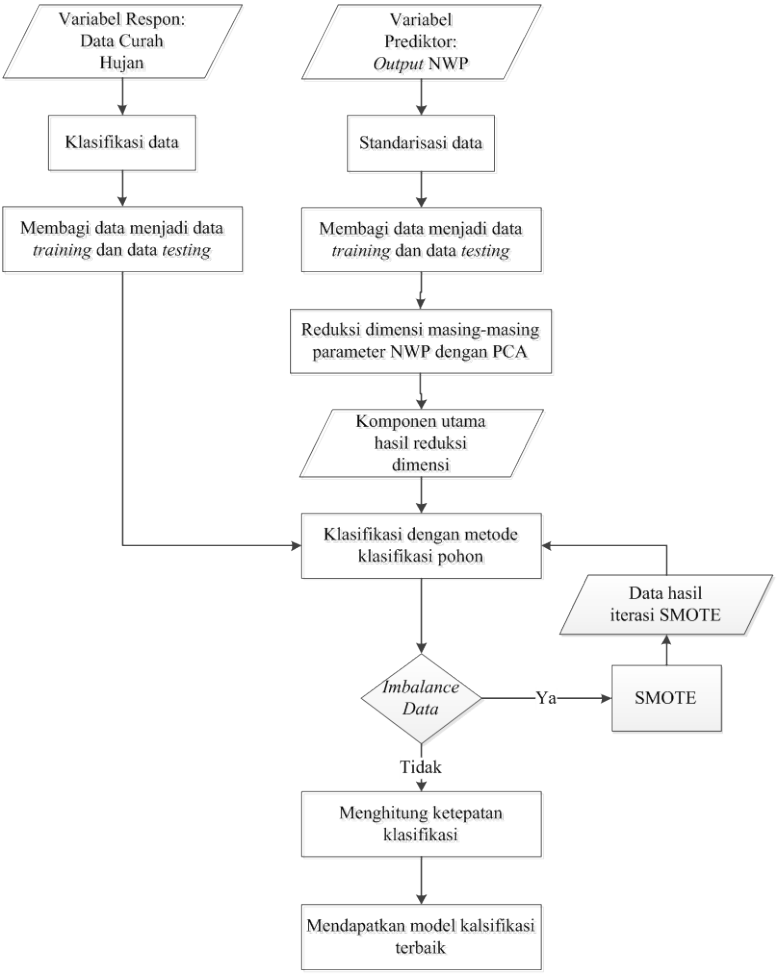
3.3 Tahapan Analisis Data

Langkah-langkah analisis data yang dilakukan dalam penelitian ini adalah sebagai berikut.

1. Melakukan standarisasi data *output* NWP dan melakukan klasifikasi pada data curah hujan sesuai dengan kategori dari BMKG.
2. Membagi data curah hujan dan *output* NWP menjadi data *training* dan data *testing*. Data *testing* diambil sebanyak 7 data terbaru dan sisanya dijadikan data *training*.

3. Mereduksi dimensi masing-masing parameter variabel *output* NWP dalam 9 grid pengukuran menggunakan *Principal Component Analysis* (PCA) dengan langkah sebagai berikut.
   a. Menghitung matriks varian kovarians
   b. Menghitung nilai *eigen value* dan *eigen vektor* dari matriks kovarians
   c. Membentuk variabel baru (komponen utama) dari *eigen vektor*
4. Melakukan klasifikasi curah hujan menggunakan metode klasifikasi pohon dimana komponen utama hasil PCA dijadikan sebagai variabel prediktor, dengan langkah sebagai berikut.
   a. Pembentukkan pohon klasifikasi
   b. Menentukan pemilah (*classifier*) menggunakan indeks gini
   c. Penentuan *node* terminal
   d. Melakukan proses SMOTE ketika terindikasi adanya *imbalance* data
   e. Mengulangi proses a sampai c dengan data hasil proses SMOTE.
   f. Penandaan label kelas
   g. Pemangkasan pohon klasifikasi
   h. Penentuan pohon klasifikasi optimal
5. Melakukan validasi menggunakan data *testing* dengan cara memasukkan data *testing* ke model pohon optimal yang telah terbentuk
6. Menghitung ketepatan klasifikasi hasil klasifikasi pohon untuk setiap wilayah pengamatan
7. Mendapatkan model kalsifikasi terbaik

Langkah pengolahan dan analisis data yang dilakukan dalam penelitian ini disajikan dengan diagram alir pada Gambar 3.2.



Gambar 3.2 Diagram Alir Analisis Data

(Halaman ini sengaja dikosongkan)

BAB IV

HASIL DAN PEMBAHASAN

BAB IV
HASIL DAN PEMBAHASAN

Bab ini membahas penyusunan MOS dengan metode klasifikasi pohon dan dilakukan validasi model klasifikasi pohon dengan menghitung ketepatan klasifikasi menggunakan nilai APER. Bagian awal disajikan deskripsi curah hujan di tiga stasiun pengamatan dan membahas reduksi dimensi data NWP menggunakan metode PCA.

4.1 Deskripsi Curah Hujan dan *Output* NWP di Wilayah Penelitian

Curah hujan dikategorikan menjadi 5 yakni cerah berawan (curah hujan ≤ 0.1 mm/hari), hujan ringan (curah hujan ≤ 20 mm/hari), hujan sedang (curah hujan ≤ 50 mm/hari), hujan lebat (curah hujan ≤ 100 mm/hari), dan hujan lebat sekali (curah hujan > 100 mm/hari) (BMKG,2006). Berdasarkan kriteria tersebut, diperoleh deskripsi curah hujan untuk masing-masing stasiun pengamatan seperti ditunjukkan pada Tabel 4.1

Tabel 4.1 Persentase Kejadian Hujan Menurut Stasiun Pengamatan

Stasiun Pengamatan	Kategori Kejadian Hujan (%)					
	Cerah Berawan	Hujan Ringan	Hujan Sedang	Hujan Lebat	Hujan Lebat Sekali	Total
Citeko	1,1	73,0	20,4	5,0	0,4	100
Kemayoran	0,7	76,3	16,0	5,7	1,3	100
Pnd. Betung	2,3	78,0	14,8	3,8	1,0	100

Tabel 4.1 menunjukkan bahwa hujan ringan sering terjadi pada ketiga stasiun pengamatan. Hujan lebat sekali jarang terjadi pada 2 stasiun pengamatan yakni Citeko dan Pondok Betung dengan persentase dibawah 1%. Sedangkan pada stasiun pengamatan Kemayoran jarang terjadi kejadian cerah berawan.

Sebelum melakukan *pre-processing* data, perlu dilakukan standarisasi pada data NWP. Hal ini dikarenakan adanya perbedaan satuan pengukuran pada masing-masing variabel

NWP. Standarisasi dilakukan dengan cara mengurangi data dengan rata-rata kemudian dibagi dengan variannya. Setelah data terstandarisasi, kemudian data dibagi menjadi data *training* dan data *testing*. Pada penelitian ini menggunakan data *testing* sebanyak 7 hari terbaru, kemudian sisanya dijadikan data *training*. Data *training* digunakan untuk membangun model sedangkan data *testing* digunakan untuk validasi model yang terbentuk.

4.2 Reduksi Dimensi Data NWP dengan Metode PCA

Kriteria penentuan variabel baru dari reduksi dimensi menggunakan metode PCA yakni berdasarkan besar proporsi keragaman yang dapat dijelaskan oleh komponen terbentuk diatas 85 persen. *Eigenvalue* dan keragaman kumulatif variabel pblh hasil reduksi PCA pada stasiun pengamatan Citeko ditampilkan pada Tabel 4.2 berikut.

Tabel 4.2 *Eigenvalue* dan Kumulatif Keragaman Variabel pblh

PC	*Eigenvalue*	Keragaman yang dijelaskan	Keragaman Kumulatif
1	8,127	0,903	0,903
2	0,431	0,048	0,951
3	0,238	0,026	0,977
4	0,104	0,012	0,989
5	0,044	0,005	0,994
6	0,028	0,003	0,997
7	0,014	0,002	0,999
8	0,008	0,001	1
9	0,001	0,000	1

Keragaman variabel pblh yang dijelaskan oleh komponen (PC) pertama sebesar 90,3 persen. Sehingga keragaman variabel pblh dapat dijelaskan dengan satu komponen. Jumlah komponen dengan kumulatif keragaman diatas 85 persen yang terbentuk dari data NWP secara lengkap ditunjukkan pada Tabel 4.3 berikut.

Tabel 4.3 *Eigenvalue* dan Keragaman PC Variabel NWP Stasiun Citeko

Variabel	Citeko		
	Jmlh PC	***Eigenvalue***	**Keragaman**
dpsdt	1	8,998	100%
mixr1	1	7,752	86,1%
mixr2	1	8,301	92,2%
mixr4	1	8,623	95,8%
omega1	2	7,055 ; 0,971	89,2%
omega2	2	6,904 ; 0,868	86,4%
omega4	2	7,281 ; 0,929	91,2%
pblh	1	8,127	90,3%
ps	1	8,976	99,7%
psl	1	8,996	100%
qgscrn	2	7,234 ; 0,790	89,2%
rh1	1	7,791	86,6%
rh2	1	8,320	92,4%
rh4	1	8,656	96,2%
rnd	1	7,915	88,0%
temp1	1	8,473	94,1%
temp2	1	8,642	96,0%
temp4	1	8,878	98,6%
tmaxscr	1	8,807	97,9%
tminscr	1	8,387	93,2%
tpan	1	8,655	96,2%
tscrn	1	8,472	94,1%
u1	1	8,477	94,2%
u2	1	8,660	96,2%
u4	1	8,920	99,1%
ustar	2	7,010 ; 1,052	89,6%

Tabel 4.3 (Lanjutan) *Eigenvalue* dan Keragaman PC Variabel NWP Stasiun Citeko

Variabel	Citeko		
	Jmlh PC	***Eigenvalue***	**Keragaman**
v1	2	6,943 ; 1,104	89,5%
v2	2	6,896 ; 1,276	90,8%
v4	1	8,720	96,9%
zg1	2	7,187 ; 0,997	91,0%
zg2	2	5,777 ; 2,773	95,0%
zg4	2	7,126 ; 1,746	98,6%

Tabel 4.3 menunjukkan bahwa jumlah komponen utama yang terbentuk dari seluruh variabel NWP di stasiun Citeko sebanyak 42 komponen. Dimana hasil reduksi untuk masing-masing variabel data NWP menghasilkan rata-rata sebanyak 1 hingga 2 komponen yang mampu menjelaskan keragaman masing-masing variabel.

Tabel 4.4 *Eigenvalue* dan Keragaman PC Variabel NWP Stasiun Kemayoran

Variabel	Kemayoran		
	Jmlh PC	***Eigenvalue***	**Keragaman**
dpsdt	1	8,998	100%
mixr1	1	8,105	90,10%
mixr2	1	8,376	93,10%
mixr4	1	8,619	95,80%
omega1	1	8,834	98,20%
omega2	1	8,155	90,60%
omega4	1	8,028	89,20%
pblh	1	8,223	91,40%
ps	1	8,997	100%
psl	1	8,998	100%

Tabel 4.4 (Lanjutan) *Eigenvalue* dan Keragaman PC Variabel NWP Stasiun Kemayoran

Variabel	Kemayoran		
	Jmlh PC	***Eigenvalue***	**Keragaman**
qgscrn	1	8,022	89,10%
rh1	1	7,976	88,60%
rh2	1	8,432	93,70%
rh4	1	8,662	96,20%
rnd	1	7,701	85,60%
temp1	1	8,570	95,20%
temp2	1	8,756	97,30%
temp4	1	8,919	99,10%
tmaxscr	1	8,775	97,50%
tminscr	1	8,306	92,30%
tpan	1	8,692	96,60%
tscrn	1	8,566	95,20%
u1	1	8,786	97,60%
u2	1	8,871	98,60%
u4	1	8,952	99,50%
ustar	1	8,188	91,00%
v1	1	8,203	91,10%
v2	1	8,322	92,50%
v4	1	8,882	98,70%
zg1	4	3,53; 2,13 ; 1,13 ; 0,94	86,20%
zg2	2	6,887 ; 0,872	86,20%
zg4	1	8,785	97,60%

Hasil reduksi dimensi variabel NWP pada stasiun Kemayoran ditampilkan pada Tabel 4.4, dimana total komponen utama yang terbentuk sebanyak 36 PC. Rata-rata variabel menghasilkan 1 komponen utama, kecuali pada variabel zg1 dan

zg2 yakni 4 dan 3 komponen dengan keragaman kumulatif 86,20%. Jadi terdapat 36 variabel prediktor untuk membangun model klasifikasi pohon pada stasiun Kemayoran.

Tabel 4.5 *Eigenvalue* dan Keragaman PC Variabel NWP Stasiun Pondok Betung

Variabel	**Pondok Betung**		
	Jmlh PC	***Eigenvalue***	**Keragaman Kumulatif**
dpsdt	1	8,998	100%
mixr1	1	8,327	92,50%
mixr2	1	8,650	96,10%
mixr4	1	8,754	97,30%
omega1	1	8,910	99,00%
omega2	1	8,556	95,10%
omega4	1	8,107	90,10%
pblh	1	8,398	93,30%
ps	1	8,998	100%
psl	1	8,998	100%
qgscrn	1	8,296	92,20%
rh1	1	8,289	92,10%
rh2	1	8,640	96,00%
rh4	1	8,762	97,40%
rnd	1	8,054	89,50%
temp1	1	8,684	96,50%
temp2	1	8,806	97,80%
temp4	1	8,925	99,20%
tmaxscr	1	8,884	98,70%
tminscr	1	8,588	95,40%
tpan	1	8,805	97,80%
tscrn	1	8,673	96,40%
u1	1	8,804	97,80%

Tabel 4.5 (Lanjutan) *Eigenvalue* dan Keragaman PC Variabel NWP Stasiun Pondok Betung

Variabel	Pondok Betung		
	Jmlh PC	***Eigenvalue***	**Keragaman**
u2	1	8,868	98,50%
u4	1	8,957	99,50%
ustar	1	8,396	93,30%
v1	1	8,429	93,70%
v2	1	8,433	93,70%
v4	1	8,892	98,80%
zg1	5	3,3; 2,24; 1,2 ; 0,83; 0,6	91,00%
zg2	2	7,367 ; 0,759	90,30%
zg4	1	8,838	98,20%

Tabel 4.5 menampilkan komponen utama yang terbentuk dari hasil reduksi dimensi pada stasiun Pondok Betung. Dari Tabel 4.5, dapat diketahui bahwa komponen utama yang terbentuk pada stasiun ini sebanyak 37 komponen. Sebagian besar variabel menghasilkan 1 komponen utama kecuali untuk variabel zg1 dan zg2 yakni 5 dan 2 komponen. Sebanyak 37 komponen tersebut akan digunakan sebagai variabel prediktor untuk membangun model klasifikasi pohon pada stasiun Pondok Betung.

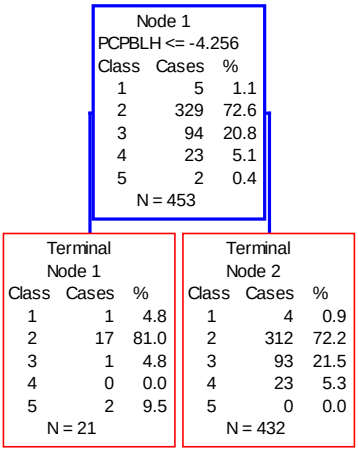
4.3 Klasifikasi Curah Hujan

Berdasarkan pada tujuan penelitian, maka dilakukan analisis klasifikasi curah hujan dengan menggunakan pendekatan *Classification and Regression Tree* (CART). Adapun variabel respon yang digunakan pada penelitian ini berupa data kategorik yaitu cerah berawan, hujan ringan, hujan sedang, hujan lebat, dan hujan lebat sekali. Sehingga pendekatan CART akan menghasilkan suatu pohon klasifikasi (*classification tree*).

Sesuai dengan prosedur algoritma CART yang telah dijelaskan pada bab tinjauan pustaka, maka tahapan pertama yang dilakukan adalah pembentukan pohon klasifikasi. Metode pemilihan pemilah pada pembentukan pohon klasifikasi menggunakan *10-vold cross validation* karena jumlah data penelitian kurang dari 3000 data.

4.3.1 Klasifikasi Curah Hujan Stasiun Citeko

Pada stasiun pengamatan Citeko, terdapat 453 data pengamatan curah hujan. Kemudian data tersebut digunakan sebagai data untuk membangun model klasifikasi pohon. Dari split plot pohon optimal pada Gambar 4.1, dapat dilihat bahwa pada terminal *node* 1 dan 2 didominasi oleh kelas 2 dengan persentase diatas 70 persen. Sehingga klasifikasi yang dihasilkan cenderung kepada kelas 2 dan menghasilkan tingkat ketepatan klasifikasi (1-APER) yang rendah yakni 7,95% pada data *training*. Sedangkan data *testing* menghasilkan ketepatan klasifikasi 100%, hal ini dikarenakan pohon optimal yang terbentuk cenderung pada kelas 2 dengan variabel PCpblh sebagai variabel pemilah. Ketika nilai variabel PCpblh $\leq$ -4,256 ataupun > -4,256, maka data *testing* akan tetap diklasifikasikan pada kelas 2. Padahal ke-7 data *testing* yang digunakan memiliki klasifikasi aktual berada pada kelas 2. Sehingga jika data *testing* dimasukkan dalam pohon optimal yang terbentuk, maka seluruh data *testing* diklasifikasikan dalam kelas 2 dan menyebabkan nilai 1-APER menjadi 100%. Secara angka, nilai 1-APER 100% memang bagus. Namun jika dilihat struktur pohon yang terbentuk maka dapat dikatakan bahwa pohon tersebut tidak bagus.



Gambar 4.1 Spliplot Pohon Optimal Stasiun Citeko Sebelum SMOTE

Dari analisa yang telah dilakukan, dapat diketahui bahwa terjadi *imbalance data* pada stasiun pengamatan Citeko, dimana jumlah suatu kelas mayor jauh lebih besar dari jumlah kelas yang lain (kelas minor). Jika dilakukan pembentukan pohon dengan kondisi jumlah kelas yang *imbalance*, akan mengakibatkan klasifikasi yang cenderung kepada kelas mayor dan mengabaikan kelas minor sehingga akurasi kelas minor sangat kecil. Kasus *imbalance data* merupakan permasalahan yang sering dijumpai dalam pengklasifikasian. Untuk menyeimbangkan jumlah data kelas minor, perlu dilakukan pra-pemrosesan menggunakan metode SMOTE (*Synthetic Minority Oversampling Technique*). SMOTE merupakan salah satu metode *oversampling* yaitu teknik pengambilan sampel untuk meningkatkan jumlah data pada kelas minor dengan cara mereplikasi jumlah data pada kelas minor secara acak sehingga jumlahnya sama dengan data pada kelas mayor. Setelah dilakukan SMOTE pada data *training* dengan iterasi sebanyak 11 kali, jumlah data pengamatan untuk stasiun Citeko menjadi 659 data.

Kemudian seluruh 659 data tersebut dijadikan sebagai data *learning* untuk membangun model klasifikasi pohon. Sedangkan

untuk data *testing* menggunakan 7 data *testing* sebelum SMOTE. Berikut penjelasan untuk masing-masing tahapan analisis klasifikasi pohon pada stasiun pengamatan Citeko dengan menggunakan kombinasi data *learning* dan *testing* tersebut.

4.3.1.1 Pembentukan Pohon Klasifikasi Maksimal

Tahapan awal yang dilakukan untuk membentuk pohon klasifikasi adalah dengan menentukan variabel pemilah.Variabel pemilah dipilih dari beberapa kemungkinan pemilah setiap variabel prediktor.

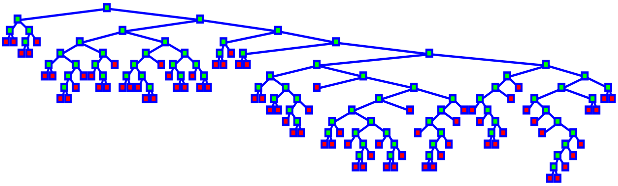
Selanjutnya dihitung Indeks Gini yang merupakan ukuran keheterogenan *node*. Indeks Gini lebih sering digunakan karena alasan kesederhanaan dalam proses perhitungan. Cara kerja Indeks Gini adalah melakukan pemilihan *node* dengan berfokus pada masing-masing *node* kanan atau kiri. Hasil perhitungan Indeks Gini kemudian digunakan utuk menentukan *goodness of split* dari masing-masing pemilah. Pemilah yang terpilih adalah variabel pemilah dan nilai variabel (*threshold*) yang memiliki nilai *goodness of split* tertinggi. Pemilah yang terpilih merupakan variabel yang terpenting dalam klasifikasi data pengamatan. Besarnya kontribusi variabel sebagai pemilah baik pemilah utama maupun pengganti pada pohon klasifikasi maksimal yang terbentuk ditunjukkan melalui suatu angka skor yang ditampilkan secara lengkap pada Lampiran 8.

Berdasarkan Lampiran 8 diperoleh informasi bahwa semua variabel prediktor menjadi pembangun dalam pembentukan pohon klasifikasi maksimal. Akan tetapi berdasarkan skor yang dihasilkan, variabel PCpblh mempunyai skor tertinggi seperti ditampilkan pada Tabel 4.6. Sehingga variabel PCpblh merupakan variabel terpenting dan menjadi pemilah utama dalam klasifikasi curah hujan di Stamet Citeko. Selain itu, terdapat beberapa variabel yang berpengaruh besar yakni PC2qgscr, PCdpsdt, PCrh2 dan PCmixr2. Sedangkan variabel lain memiliki skor di bawah 50.

Tabel 4.6 Variabel Penting Pembentukan Pohon Klasifikasi Maksimal Stasiun Citeko Setelah SMOTE

Variabel	Skor Variabel	
PCpblh	100	`||||||||||||||||||||||||||||||||||||||||||`
PC2qgscr	97,43	`|||||||||||||||||||||||||||||||||||||||||`
PCdpsdt	78,12	`||||||||||||||||||||||||||||||||`
PCrh2	68,89	`|||||||||||||||||||||||||`
PCmixr2	67,05	`||||||||||||||||||||||||`

Hasil penyekatan rekursif biner dari data pengamatan yang digunakan akan menghasilkan pohon klasifikasi yang berukuran relatif besar dengan tingkat kedalaman yang tinggi. Pohon tersebut merupakan pohon klasfikasi maksimal yang ditampilkan pada Gambar 4.2 dengan *node* sebanyak 79 dan kedalaman 15 tingkatan.

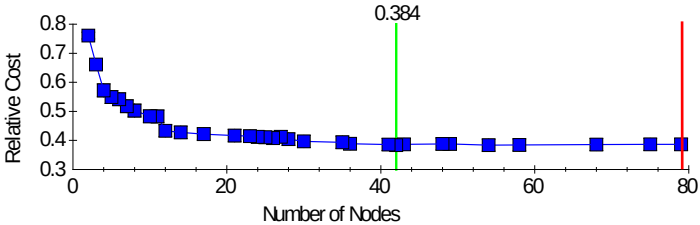


Gambar 4.2 Topologi Pohon Klasifikasi Maksimal untuk Curah Hujan Stasiun Citeko Setelah SMOTE

4.3.1.2 Pemangkasan Pohon Klasifikasi Maksimal (*Prunning*)

Pohon yang besar dan kompleks akan mempersulit peneliti dalam hal interpretasi hasil klasifikasi. Untuk mempermudah proses analisis, maka dilakukan pemangkasan secara iteratif terhadap pohon klasifikasi maksimal yang terbentuk berdasarkan kriteria *cross-validated relative cost*. Setiap hasil pemangkasan memiliki nilai *relative cost* tertentu, kemudian dipilih hasil pemangkasan dengan nilai *relative cost* yang minimum.

Gambar 4.3 menampilkan adanya perbedaan nilai *relative cost* yang dihasilkan oleh pohon klasifikasi maksimal dengan pohon klasifikasi yang dianggap optimal. Pohon klasifikasi maksimal ditunjukkan oleh garis berwarna merah sedangkan pohon klasifikasi optimal ditunjukkan oleh garis berwarna hijau.



Gambar 4.3 Plot *Relative Cost* Klasifikasi Curah Hujan Stasiun Citeko Setelah SMOTE

Berdasarkan Gambar 4.3, pohon klasifikasi maksimal yang terbentuk terdiri dari 79 *terminal nodes* dan *relative cost* sebesar 0,386 ± 0,0211 yang dapat dilihat pada Tabel 4.7. Pemangkasan pohon dilakukan secara iteratif berdasarkan *cross validated relative cost* yang minimum. Tabel 4.7 menunjukkan bahwa nilai *cross validated relative cost* yang minimum adalah pada saat *terminal nodes* sebanyak 42. Sehingga dapat dikatakan bahwa pohon klasifikasi optimal yang terbentuk terdiri dari 42 *terminal nodes*. Karena nilai *relative cost* pohon klasifikasi optimal lebih kecil maka pohon klasifikasi optimal dipilih sebagai pohon yang layak untuk pohon klasifikasi curah hujan pada stasiun pengamatan Citeko.

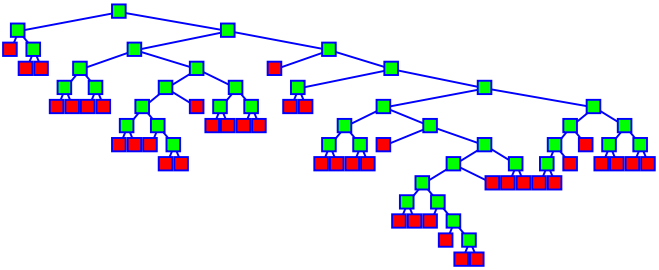
Tabel 4.7 Pembentukan Pohon Klasifikasi Stasiun Citeko Setelah SMOTE

Tree Number	***Terminal Nodes***	***Cross-validated Relative Cost***	***Resubsitution Relative Cost***
1	79	0,386 ± 0,021	0,053
9*	42	0,384 ± 0,021	0,118
24	11	0,482 ± 0,022	0,338
25	10	0,483 ± 0,022	0,357
26	8	0,503 ± 0,022	0,398
27	7	0,518 ± 0,022	0,426
28	6	0,541 ± 0,022	0,456
29	5	0,549 ± 0,023	0,494
30	4	0,572 ± 0,020	0,537
31	3	0,661 ± 0,012	0,595
32	2	0,761 ± 0,011	0,750

*Pohon Klasifikasi Optimal

4.3.1.3 Pemilihan Pohon Klasifikasi Optimal

Hasil pemangkasan pohon maksimal secara iteratif menghasilkan pohon klasifikasi optimal dengan jumlah *terminal nodes* sebanyak 42 *node* ditampilkan pada Gambar 4.4. Nilai *cross validatied relative cost* pohon optimal yaitu sebesar 0,384 ± 0,021 yang berarti nilai kesalahan prediksi besarnya curah hujan dari klasifikasi pohon maksimal berkisar antara 0,405 sampai 0,363 dengan *resubstition relative cost* sebesar 0,053.



Gambar 4.4 Topologi Pohon Klasifikasi Optimal untuk Klasifikasi Curah Hujan pada Stasiun Citeko Setelah SMOTE

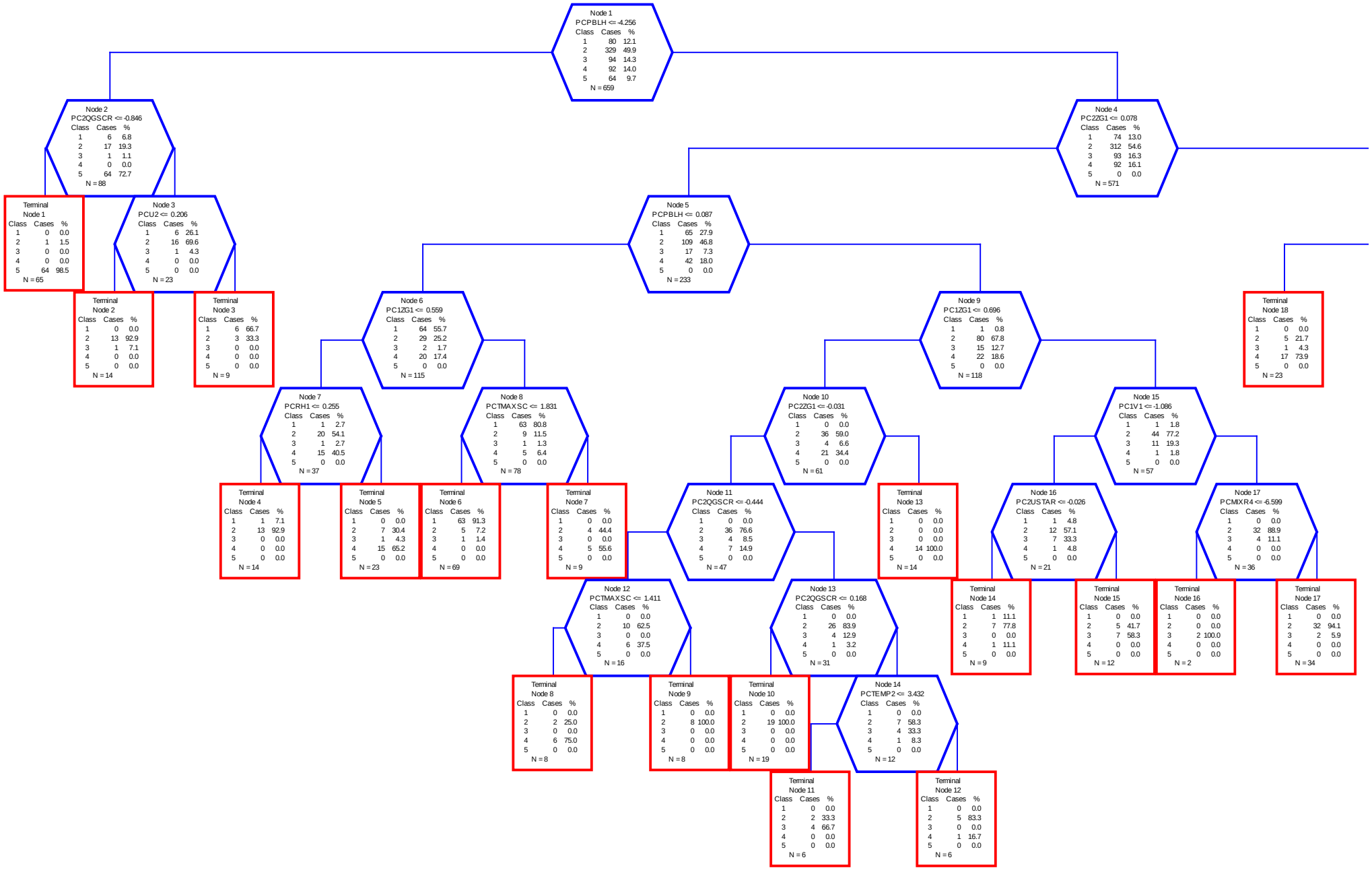
Berdasarkan topologi pohon klasifikasi optimal, diketahui bahwa PCpblh merupakan variabel pemilah yang utama dan paling penting dalam menentukan klasifikasi curah hujan di stasiun pengamatan Citeko. Pada Tabel 4.8, skor variabel PCpblh sebesar 100 karena mampu memberikan nilai penurunan keheterogenan tertinggi pada *node* utama. Selain itu ada 37 variabel lain yang juga berkontribusi dalam pembentukan pohon klasifikasi optimal, hasil selengkapnya disajikan dalam Lampiran 9.

Tabel 4.8 Variabel Penting Pembentukan Pohon Klasifikasi Optimal Stasiun Citeko Setelah SMOTE

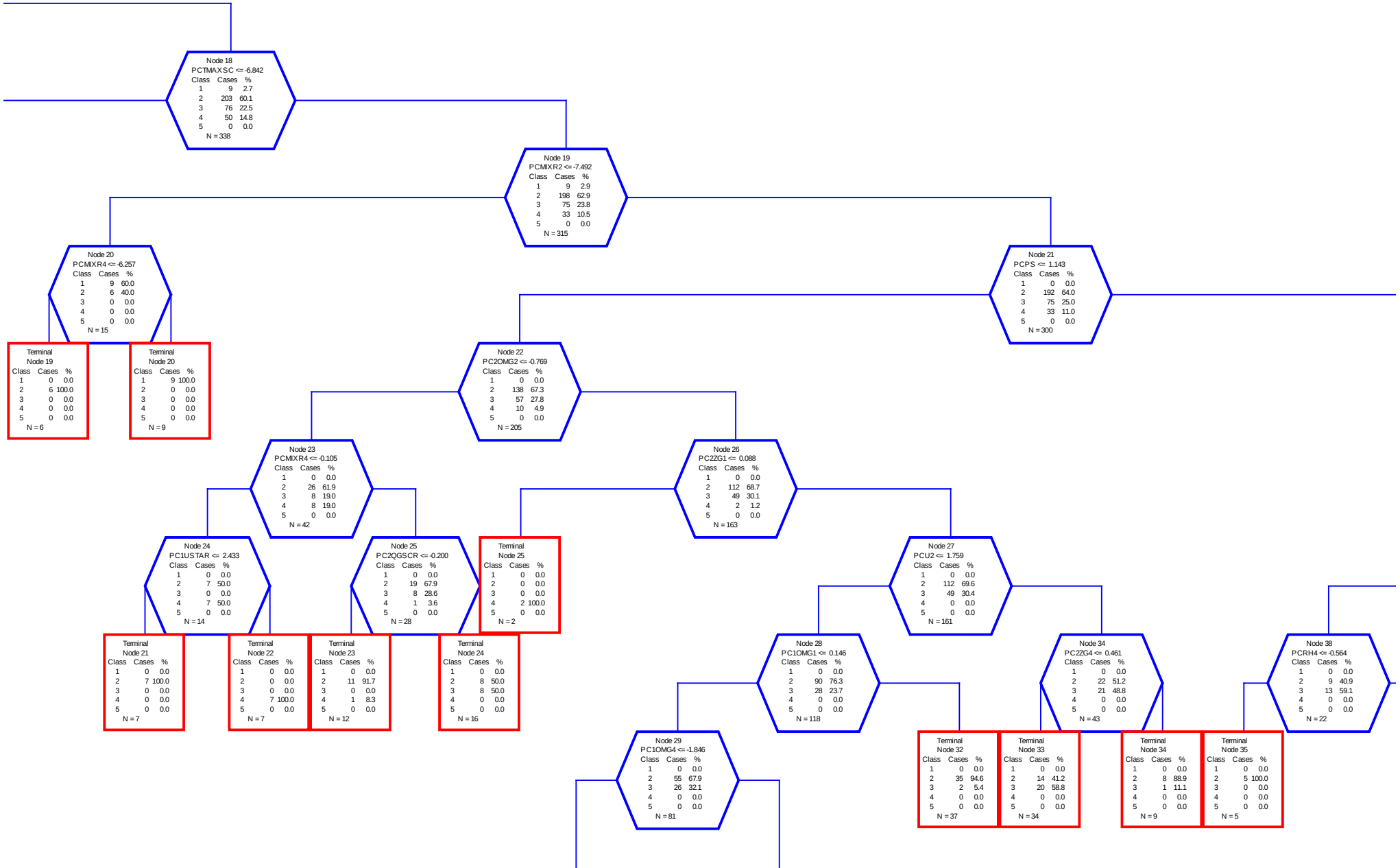
Variabel	Skor Variabel
PCpblh	100
PC2qgscr	96,91
PCdpsdt	75,01
PCrh2	69,12
PCmixr2	64,14

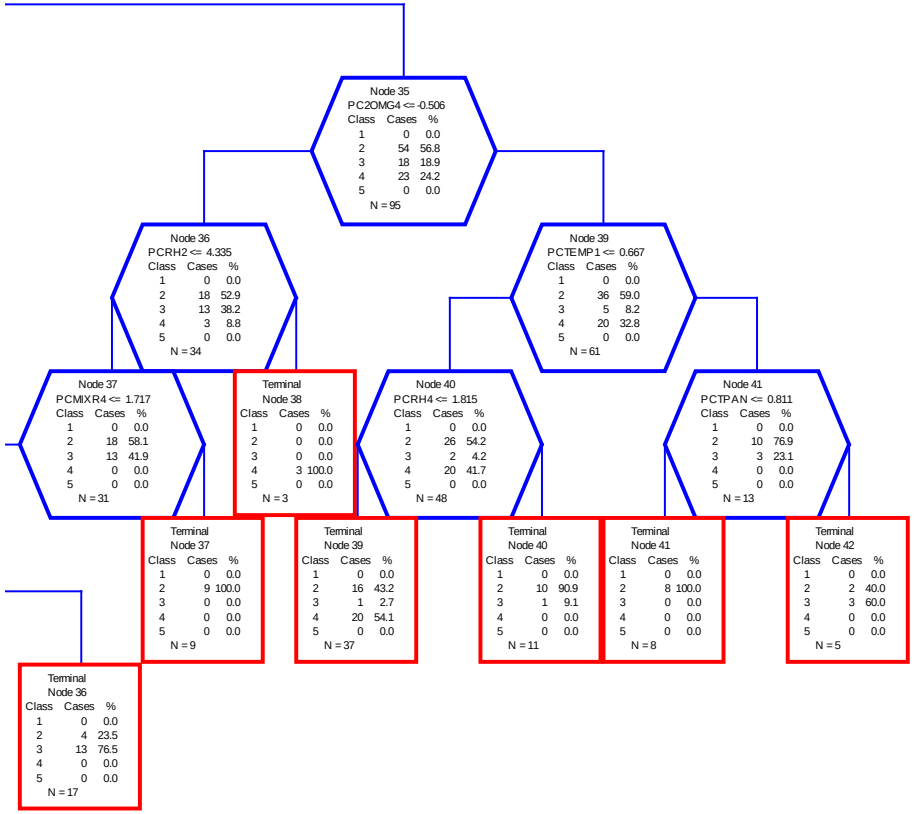
Variabel utama PCpblh memilah *node* utama menjadi *node* kanan dan kiri dengan ketentuan nilai PCpblh ≤ 4,256 akan dipilah menjadi *node* kiri. Sedangkan jika nilai PCpblh > 4,256 akan dipilah menjadi *node* kanan. Gambar 4.5 merupakan visualisasi struktur pohon klasifikasi optimal.

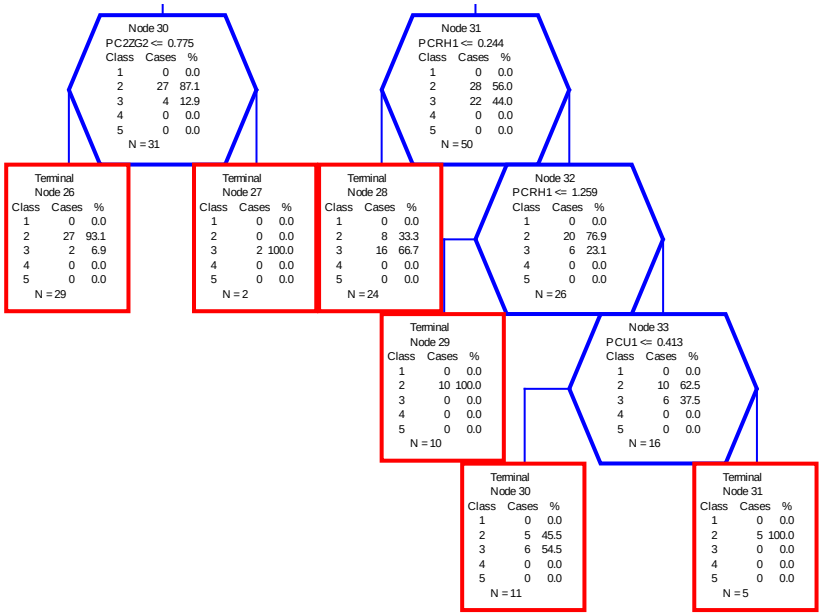
Suatu *node* akan terus dipilah menjadi *node* anak baru (kiri dan kanan) sesuai prosedur *binary recursive partitioning*, sampai *node* tersebut telah dianggap memiliki anggota yang homogen atau jika *node* tersebut hanya memiliki 1 anggota pengamatan maka *node* akan menjadi *node* terminal dan tidak akan dipilah lagi. Pohon klasifikasi optimal yang terbentuk terdiri atas 42 *node* terminal seperti pada Gambar 4.5. Masing-masing *node* terminal tersebut memiliki karakteristik tertentu dan diprediksi sebagai kelas variabel respon tertentu sesuai dengan label kelas yang diberikan. Berdasarkan hasil penelusuran 42 terminal *node* tersebut, Tabel 4.9 menampilkan rangkuman pengklasifikasian curah hujan menurut indikasi kesamaan label kelas setiap *node* terminal.



Gambar 4.5 Split Plot Pohon Optimal Stasiun Citeko Setelah SMOTE


Node 18
PCTMAXSC <= -6.842
Class Cases %
1 9 2.7
2 203 60.1
3 76 22.5
4 50 14.8
5 0 0.0
N = 338
Node 19
PCMIXR2 <= -7.492
Class Cases %
1 9 2.9
2 198 62.9
3 75 23.8
4 33 10.5
5 0 0.0
N = 315
Node 20
PCMIXR4 <= -6.257
Class Cases %
1 9 60.0
2 6 40.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 15
Terminal
Node 19
Class Cases %
1 0 0.0
2 6 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 6
Terminal
Node 20
Class Cases %
1 9 100.0
2 0 0.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 9
Node 21
PCPS <= 1.143
Class Cases %
1 0 0.0
2 192 64.0
3 75 25.0
4 33 11.0
5 0 0.0
N = 300
Node 22
PC2OMG2 <= -0.769
Class Cases %
1 0 0.0
2 138 67.3
3 57 27.8
4 10 4.9
5 0 0.0
N = 205
Node 23
PCMIXR4 <= -0.105
Class Cases %
1 0 0.0
2 26 61.9
3 8 19.0
4 8 19.0
5 0 0.0
N = 42
Node 26
PC22G1 <= 0.088
Class Cases %
1 0 0.0
2 112 68.7
3 49 30.1
4 2 1.2
5 0 0.0
N = 163
Node 24
PC1USTAR <= 2.433
Class Cases %
1 0 0.0
2 7 50.0
3 0 0.0
4 7 50.0
5 0 0.0
N = 14
Node 25
PC2QGSCR <= -0.200
Class Cases %
1 0 0.0
2 19 67.9
3 8 28.6
4 1 3.6
5 0 0.0
N = 28
Terminal
Node 25
Class Cases %
1 0 0.0
2 0 0.0
3 0 0.0
4 2 100.0
5 0 0.0
N = 2
Node 27
PCU2 <= 1.759
Class Cases %
1 0 0.0
2 112 69.6
3 49 30.4
4 0 0.0
5 0 0.0
N = 161
Terminal
Node 21
Class Cases %
1 0 0.0
2 7 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 7
Terminal
Node 22
Class Cases %
1 0 0.0
2 0 0.0
3 0 0.0
4 7 100.0
5 0 0.0
N = 7
Terminal
Node 23
Class Cases %
1 0 0.0
2 11 91.7
3 0 0.0
4 1 8.3
5 0 0.0
N = 12
Terminal
Node 24
Class Cases %
1 0 0.0
2 8 50.0
3 8 50.0
4 0 0.0
5 0 0.0
N = 16
Node 28
PC1OMG1 <= 0.146
Class Cases %
1 0 0.0
2 90 76.3
3 28 23.7
4 0 0.0
5 0 0.0
N = 118
Node 34
PC22G4 <= 0.461
Class Cases %
1 0 0.0
2 22 51.2
3 21 48.8
4 0 0.0
5 0 0.0
N = 43
Node 38
PCRH4 <= -0.564
Class Cases %
1 0 0.0
2 9 40.9
3 13 59.1
4 0 0.0
5 0 0.0
N = 22
Node 29
PC1OMG4 <= -1.846
Class Cases %
1 0 0.0
2 55 67.9
3 26 32.1
4 0 0.0
5 0 0.0
N = 81
Terminal
Node 32
Class Cases %
1 0 0.0
2 35 94.6
3 2 5.4
4 0 0.0
5 0 0.0
N = 37
Terminal
Node 33
Class Cases %
1 0 0.0
2 14 41.2
3 20 58.8
4 0 0.0
5 0 0.0
N = 34
Terminal
Node 34
Class Cases %
1 0 0.0
2 8 88.9
3 1 11.1
4 0 0.0
5 0 0.0
N = 9
Terminal
Node 35
Class Cases %
1 0 0.0
2 5 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 5


Node 35
PC2OMG4 <= -0.506
Class Cases %
1 0 0.0
2 54 56.8
3 18 18.9
4 23 24.2
5 0 0.0
N = 95
Node 36
PCRH2 <= 4.335
Class Cases %
1 0 0.0
2 18 52.9
3 13 38.2
4 3 8.8
5 0 0.0
N = 34
Node 39
PCTEMP1 <= 0.667
Class Cases %
1 0 0.0
2 36 59.0
3 5 8.2
4 20 32.8
5 0 0.0
N = 61
Node 37
PCMIXR4 <= 1.717
Class Cases %
1 0 0.0
2 18 58.1
3 13 41.9
4 0 0.0
5 0 0.0
N = 31
Terminal
Node 38
Class Cases %
1 0 0.0
2 0 0.0
3 0 0.0
4 3 100.0
5 0 0.0
N = 3
Node 40
PCRH4 <= 1.815
Class Cases %
1 0 0.0
2 26 54.2
3 2 4.2
4 20 41.7
5 0 0.0
N = 48
Node 41
PCTPAN <= 0.811
Class Cases %
1 0 0.0
2 10 76.9
3 3 23.1
4 0 0.0
5 0 0.0
N = 13
Terminal
Node 37
Class Cases %
1 0 0.0
2 9 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 9
Terminal
Node 39
Class Cases %
1 0 0.0
2 16 43.2
3 1 2.7
4 20 54.1
5 0 0.0
N = 37
Terminal
Node 40
Class Cases %
1 0 0.0
2 10 90.9
3 1 9.1
4 0 0.0
5 0 0.0
N = 11
Terminal
Node 41
Class Cases %
1 0 0.0
2 8 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 8
Terminal
Node 42
Class Cases %
1 0 0.0
2 2 40.0
3 3 60.0
4 0 0.0
5 0 0.0
N = 5
Terminal
Node 36
Class Cases %
1 0 0.0
2 4 23.5
3 13 76.5
4 0 0.0
5 0 0.0
N = 17


Node 30
PC2ZG2 <= 0.775
Class Cases %
1 0 0.0
2 27 87.1
3 4 12.9
4 0 0.0
5 0 0.0
N = 31
Node 31
PCRH1 <= 0.244
Class Cases %
1 0 0.0
2 28 56.0
3 22 44.0
4 0 0.0
5 0 0.0
N = 50
Terminal
Node 26
Class Cases %
1 0 0.0
2 27 93.1
3 2 6.9
4 0 0.0
5 0 0.0
N = 29
Terminal
Node 27
Class Cases %
1 0 0.0
2 0 0.0
3 2 100.0
4 0 0.0
5 0 0.0
N = 2
Terminal
Node 28
Class Cases %
1 0 0.0
2 8 33.3
3 16 66.7
4 0 0.0
5 0 0.0
N = 24
Node 32
PCRH1 <= 1.259
Class Cases %
1 0 0.0
2 20 76.9
3 6 23.1
4 0 0.0
5 0 0.0
N = 26
Terminal
Node 29
Class Cases %
1 0 0.0
2 10 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 10
Node 33
PCU1 <= 0.413
Class Cases %
1 0 0.0
2 10 62.5
3 6 37.5
4 0 0.0
5 0 0.0
N = 16
Terminal
Node 30
Class Cases %
1 0 0.0
2 5 45.5
3 6 54.5
4 0 0.0
5 0 0.0
N = 11
Terminal
Node 31
Class Cases %
1 0 0.0
2 5 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 5

Tabel 4.9 Kelas Curah Hujan Stasiun Citeko Setelah SMOTE pada Masing-Masing Terminal *Node*

Kelas	*Terminal Node*	Persentase	*Terminal Node*	Persentase
1	3	89,2	20	100
	6	96,8		
2	2	78,8	26	79,4
	4	76	29	100
	9	100	31	100
	10	100	32	83,3
	12	58,3	34	69,6
	14	47,7	35	100
	17	82,1	37	100
	19	100	40	74,1
	21	100	41	100
	23	75,5		
3	11	87,5	28	87,5
	15	83,1	30	80,8
	16	100	33	83,3
	24	77,8	36	91,9
	27	100	42	84
4	5	83,6	22	100
	7	81,7	25	100
	8	91,5	38	100
	13	100	39	78,6
	18	87,7		
5	1	99,7		

Dari tabel 4.9 dapat diketahui bahwa dari 42 terminal *node* yang terbentuk, kelas 2 merupakan *node* yang paling banyak terbentuk. Hal ini dikarenakan jumlah data pengamatan pada kelas 2 paling tinggi dibandingkan kelas lainnya. Secara

keseluruhan dapat diketahui bahwa terdapat 78 pengamatan dalam kelas cerah berawan, 238 pengamatan masuk dalam kelas hujan ringan, 81 pengamatan termasuk dalam kelas hujan sedang, 89 pengamatan masuk dalam kelas hujan lebat dan 64 pengamatan yang termasuk dalam kelas hujan lebat sekali.

Penelusuran struktur pohon klasifikasi optimal terhadap *node* terminal dapat memberikan informasi tentang karakteristik kelas *node* terminal dengan persentase tertinggi untuk masing-masing kelas. Karakteristik kelas curah hujan pada masing-masing *node* terminal disajikan pada Tabel 4.10.

Tabel 4.10 Karakteristik Kelas Curah Hujan Stasiun Citeko Setelah SMOTE

Kelas	**Karakteristik**
Cerah Berawan (1)	PCpblh > -4,256 PC2zg1 > 0,078 PCtmaxscr > -6,842 PCmixr2 ≤ -7,492 PCmixr4 > -6,257
Hujan Ringan (2)	PCpblh > -4,256 PC2zg1 ≤ 0,078 PCpblh > 0,087 PC1zg1 ≤ 0,696 PC2zg1 ≤ -0,031 PC2qgscr > -0,444 PC2qgscr ≤ 0,168
Hujan Sedang (3)	PCpblh > -4,256 PCtmaxscr > -6,842 PCmixr2 > -7,492 PCps ≤ 1,142 PC2omega2 > -0,769 PC2zg1 > 0,088 PCu2 ≤ 1,758 PC1omega1 ≤ 0,146 PC1omega4 ≤ -1,846 PC2zg2 > 0,775
Hujan Lebat (4)	PCpblh > -4,256 PC2zg1 ≤ 0,078 PCpblh > 0,087

Tabel 4.10 (Lanjutan) Karakteristik Kelas Curah Hujan Stasiun Citeko Setelah SMOTE

Kelas	Karakteristik
Hujan Lebat (4)	PCpblh > -4,256 PC2zg1 ≤ 0,078 PCpblh > 0,087 PC1zg1 ≤ 0,696 PC2zg1 > -0,031 PC2zg1 ≤ 0,078
Hujan Lebat Sekali (5)	PCpblh ≤ -4,256 PC2qgscr ≤ -0,845

4.3.1.4 Hasil Ketepatan Klasifikasi Klasifikasi Pohon

Tingkat keakuratan hasil klasifikasi pohon optimal yang dihasilkan dari data *learning* dapat dihitung berdasarkan Tabel 4.11.

Tabel 4.11 Klasifikasi Curah Hujan Data *Learning* pada Pohon Optimal Stasiun Citeko Sebelum SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	1	0	1	3	0	20	4
2	36	0	60	218	15	0	329
3	11	0	17	65	1	18	77
4	1	0	4	18	0	17	5
5	1	0	1	0	0	0	2

Berdasarkan Tabel 4.11, kesalahan klasifikasi kelas pengamatan terjadi pada seluruh kelas. Kesalahan klasifikasi terbesar terjadi pada kelas 2 (hujan ringan) yakni 329 kesalahan. artinya tidak ada 1 pun pengamatan kelas 2 yang diklasifikasikan dengan benar. Hal serupa juga terjadi pada pengamatan kelas 5 (hujan lebat sekali) dimana salah klasifikasi dalam kelas 1 (cerah berawan) dan kelas 3 (hujan sedang).

Menggunakan informasi pada Tabel 4.11, maka ketepatan klasifikasi data *learning* sebelum proses SMOTE dapat dihitung sebagai berikut:

$$1 - APER = \left(1 - \frac{4 + 329 + 77 + 5 + 2}{453}\right) \times 100\% = 7{,}95\%$$

Pohon klasifikasi optimal yang terbentuk perlu divalidasi untuk mengetahui apakah pohon klasifikasi tersebut layak dan dapat digunakan untuk mengklasifikasi data baru. Tingkat keakuratan hasil klasifikasi pohon optimal yang dihasilkan dari data *testing* dapat dihitung berdasarkan Tabel 4.12.

Tabel 4.12 Klasifikasi Curah Hujan Data *Testing* pada Pohon Optimal Stasiun Citeko Sebelum SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	0	0	0	0	0	0	0
2	0	7	0	0	0	100	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0

Sehingga dapat dihitung besarnya ketepatan klasifikasi untuk data *testing* sebelum proses SMOTE sebagai berikut:

$$1 - APER = \left(1 - \frac{0}{7}\right) \times 100\% = 100\ \%$$

Selanjutnya akan ditampilkan hasil klasifikasi curah hujan pada data *learning* maupun *testing*menggunakan pohon optimal setelah diproses dengan SMOTE.

Tabel 4.13 Klasifikasi Curah Hujan Data *Learning* pada Pohon Optimal Stasiun Citeko Setelah SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	73	5	0	2	0	91,25	7
2	16	191	83	34	5	58,05	138
3	4	45	29	15	1	30,85	65
4	3	15	10	64	0	69,57	28
5	0	2	0	0	62	96,88	2

Berdasarkan Tabel 4.13, kesalahan klasifikasi kelas pengamatan terjadi pada seluruh kelas. Sebanyak 7 pengamatan yang secara aktual termasuk kelas 1 (cerah berawan) namun salah diklasifikasikan sebagai sebagai kelas 2 (hujan ringan) dan kelas 4 (hujan lebat). Kemudian sebanyak 138 pengamatan yang secara

aktual termasuk kelas 2 (hujan ringan) namun salah di klasifikasikan sebagai sebagai kelas 1 (cerah berawan), 3 (hujan sedang), 4 (hujan lebat) dan 5 (hujan lebat sekali). Kesalahan klasifikasi juga terjadi pada kelas 3 (hujan sedang) dimana sebanyak 65 pengamatan berada pada kelas 1, 2, 4 dan 5. Selanjutnya sebanyak 28 pengamatan yang secara aktual masuk kelas 4 (hujan lebat), namun salah diklasifikasikan sebagai kelas 1 (cerah berawan), 2 (hujan ringan) dan 3 (hujan sedang). Sedangkan untuk kelas 5, hanya 2 pengamatan yang salah diklasifikasikan menjadi kelas 2 (hujan ringan).

Menggunakan informasi pada Tabel 4.13, maka ketepatan klasifikasi data *learning* dapat dihitung sebagai berikut:

$$1 - APER = \left(1 - \frac{7 + 138 + 65 + 28 + 2}{659}\right) \times 100\% = 63{,}58\%$$

Hasil perhitungan ketepatan klasifikasi data *learning* sebesar 63,58 persen. Artinya pohon klasifikasi optimal mampu mengklasifikasikan pengamatan curah hujan kedalam kelas kategori hujan dengan tepat sebesar 63,58 persen.

Pohon klasifikasi optimal yang terbentuk perlu divalidasi untuk mengetahui apakah pohon klasifikasi tersebut layak dan dapat digunakan untuk mengklasifikasi data baru. Tingkat keakuratan hasil klasifikasi pohon optimal yang dihasilkan dari data *testing* dapat dihitung berdasarkan Tabel 4.14.

Tabel 4.14 Klasifikasi Curah Hujan Data *Testing* pada Pohon Optimal Stasiun Citeko Setelah SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	0	0	0	0	0	0	0
2	0	2	4	1	0	28,57	5
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0

Sehingga dapat dihitung besarnya ketepatan klasifikasi untuk data *testing* sebagai berikut:

$$1 - APER = \left(1 - \frac{5}{7}\right) \times 100\% = 28{,}57\%$$

Berikut adalah perbandingan hasil ketepatan klasifikasi pohon maksimal dengan pohon optimal yang ditunjukkan oleh Tabel 4.15.

Tabel 4.15 Perbandingan Ketepatan Klasifikasi Pohon Maksimal dan Pohon Optimal Stasiun Citeko

	Pohon Klasifikasi	**Ketepatan Klasifikasi (%)**	
		Learning	***Testing* Data Baru**
Sebelum SMOTE	Pohon Maksimal	50,99	
	Pohon Optimal	7,95	100
Setelah SMOTE	Pohon Maksimal	64,95	
	Pohon Optimal	63,58	28,57

Berdasarkan Tabel 4.15, dapat diketahui bahwa setelah dilakukan SMOTE, tidak terjadi peningkatan nilai ketepatan yang signifikan pada pohon maksimal. Sebaliknya, setelah dilakukan SMOTE, peningkatan ketepatan klasifikasi sangat terlihat pada pohon optimal kecuali pada *testing* data baru. Hal ini dikarenakan sebelum dilakukan SMOTE, pohon optimal yang terbentuk hanya mengklasifikan data pada kelas 2 sehingga ketepatan yang dihasilkan mencapai 100%. Setelah dilakukan SMOTE, pohon optimal yang terbentuk mampu mengklasifikasikan data pada 5 kelas yang berbeda dengan ketepatan klasifikasi pada *testing* data baru sebesar 28,57%. Artinya, pohon optimal yang terbenuk setelah proses SMOTE, mampu mengklasifikasikan data baru dengan tepat sebesar 28,57%.

Berdasarkan Tabel 4.15, dapat diketahui bahwa secara keseluruhan ketepatan klasifikasi pohon maksimal lebih tinggi daripada pohon optimal. Hal ini dikarenakan pohon klasifikasi maksimal memiliki *node* yang paling banyak dengan melibatkan

lebih banyak variabel prediktor sebagai pemilah *node* sehingga kemungkinan klasifikasi data dengan tepat cenderung lebih besar.

4.3.2 Klasifikasi Curah Hujan Stasiun Kemayoran

Klasifikasi curah hujan di stasiun pengamatan Kemayoran dilakukan dengan menggunakan langkah yang sama seperti stasiun pengamatan Citeko. Variabel prediktor merupakan 36 komponen utama hasil dari reduksi dimensi menggunakan PCA dengan jumlah data pengamatan sebanyak 293. Kemudian data tersebut digunakan untuk membangun model klasifikasi pohon.

4.3.2.1 Pembentukan Pohon Klasifikasi Maksimal

Pemilah yang terpilih merupakan variabel yang terpenting dalam klasifikasi data pengamatan. Besarnya kontribusi variabel sebagai pemilah baik pemilah utama maupun pengganti pada pohon klasifikasi maksimal yang terbentuk ditunjukkan melalui suatu angka skor yang ditampilkan secara lengkap pada Lampiran 12.

Berdasarkan Lampiran 12, diperoleh informasi bahwa seluruh variabel prediktor menjadi pembangun dalam pembentukan pohon klasifikasi maksimal. Akan tetapi berdasarkan skor yang dihasilkan, variabel PC2zg2 mempunyai skor tertinggi seperti ditampilkan pada Tabel 4.16. Sehingga variabel PC2zg2 merupakan variabel terpenting dan menjadi pemilah utama dalam klasifikasi curah hujan di Stasiun Kemayoran.

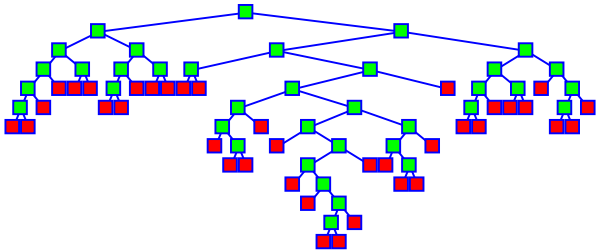
Tabel 4.16 Variabel Penting Pembentukan Pohon Klasifikasi Maksimal Stasiun Kemayoran Sebelum SMOTE

<table>
<tr><th>Variabel</th><th>Skor Variabel</th><th></th></tr>
<tr><td>PC2zg2</td><td>100</td><td>||||||||||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCrh1</td><td>95,43</td><td>||||||||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCmixr4</td><td>93,96</td><td>|||||||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCtemp2</td><td>91,32</td><td>||||||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCzg4</td><td>84,86</td><td>|||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCdpsdt</td><td>83,96</td><td>|||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCmixr2</td><td>76,21</td><td>||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCrh2</td><td>69,30</td><td>|||||||||||||||||||||||||||||</td></tr>
<tr><td>PCqgscrn</td><td>62,53</td><td>||||||||||||||||||||||||||</td></tr>
</table>

Tabel 4.16 (Lanjutan) Variabel Penting Pembentukan Pohon Klasifikasi Maksimal Stasiun Kemayoran Sebelum SMOTE

Variabel	Skor Variabel	
PCv4	58,27	|||||||||||||||||||||||||||
PCtemp4	56,11	|||||||||||||||||||||||||
PCpblh	54,65	|||||||||||||||||||||||||
PCtmaxsc	53,91	|||||||||||||||||||||||

Hasil penyekatan rekursif biner dari data pengamatan yang digunakan menghasilkan pohon klasifikasi yang berukuran relatif besar dengan tingkat kedalaman yang tinggi. Pohon tersebut merupakan pohon klasfikasi maksimal yang ditampilkan pada Gambar 4.6 dengan terminal *node* sebanyak 38 dan kedalaman 12 tingkatan.

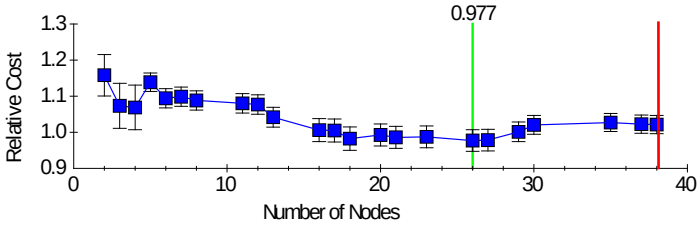


Gambar 4.6 Topologi Pohon Klasifikasi Maksimal untuk Klasifikasi Curah Hujan pada Stasiun Kemayoran Sebelum SMOTE

4.3.2.2 Pemangkasan Pohon Klasifikasi Maksimal (*Prunning*)

Untuk mempermudah proses analisis, maka dilakukan pemangkasan secara iteratif terhadap pohon klasifikasi maksimal yang terbentuk berdasarkan kriteria *cross-validated relative cost*. Setiap hasil pemangkasan memiliki nilai *relative cost* tertentu, kemudian dipilih hasil pemangkasan dengan nilai *relative cost* yang minimum.

Gambar 4.7 menampilkan adanya perbedaan nilai *relative cost* yang dihasilkan oleh pohon klasifikasi maksimal dengan pohon klasifikasi yang dianggap optimal. Pohon klasifikasi maksimal ditunjukkan oleh garis berwarna merah sedangkan pohon klasifikasi optimal ditunjukkan oleh garis berwarna hijau.



Gambar 4.7 Plot *Relative Cost* Klasifikasi Curah Hujan Stasiun Kemayoran Sebelum SMOTE

Berdasarkan Gambar 4.7, pohon klasifikasi maksimal yang terbentuk terdiri dari 38 *terminal nodes* dan *relative cost* sebesar 1,021 ± 0,025 yang dapat dilihat pada Tabel 4.17. Pemangkasan pohon dilakukan secara iteratif berdasarkan *cross validated relative cost* yang minimum. Tabel 4.17 menunjukkan bahwa nilai *cross validated relative cost* yang minimum adalah pada saat *terminal nodes* sebanyak 26. Sehingga dapat dikatakan bahwa pohon klasifikasi optimal yang terbentuk terdiri dari 26 *terminal nodes*. Karena nilai *relative cost* pohon klasifikasi optimal lebih kecil maka pohon klasifikasi optimal dipilih sebagai pohon yang layak untuk pohon klasifikasi curah hujan pada stasiun pengamatan Kemayoran.

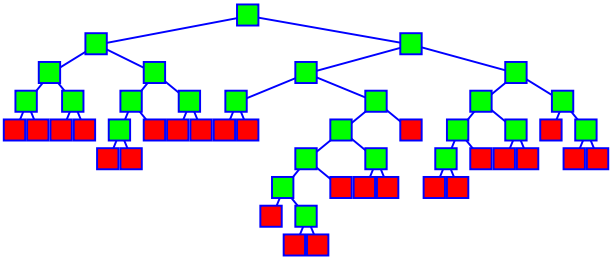
Tabel 4.17 Pembentukan Pohon Klasifikasi Stasiun Kemayoran Sebelum SMOTE

Tree Number	***Terminal Nodes***	***Cross-validated Relative Cost***	***Resubsitution Relative Cost***
1	38	1,021 ± 0.025	0,083
7*	26	0,977 ± 0.030	0,128
15	12	1,077 ± 0.027	0,306
16	11	1,080 ± 0.027	0,332
17	8	1,080 ± 0.027	0,415
18	7	1,099 ± 0.026	0,444
19	6	1,094 ± 0.027	0,517
20	5	1,139 ± 0.026	0,561
21	4	1,069 ± 0.062	0,615
22	3	1,073 ± 0.062	0,750

*Pohon Klasifikasi Optimal

4.3.2.3 Pemilihan Pohon Klasifikasi Optimal

Hasil pemangkasan pohon maksimal secara iteratif menghasilkan pohon klasifikasi optimal dengan jumlah *terminal nodes* sebanyak 26 *node* ditampilkan pada Gambar 4.8. Nilai *cross validatied relative cost* pohon optimal yaitu sebesar 0,977 ± 0,030 yang berarti nilai kesalahan prediksi besarnya curah hujan dari klasifikasi pohon maksimal berkisar antara 0,947 sampai 1,007 dengan *resubstition relative cost* sebesar 0,128.



Gambar 4.8 Topologi Pohon Klasifikasi Optimal untuk Klasifikasi Curah Hujan pada Stasiun Kemayoran Sebelum SMOTE

Berdasarkan topologi pohon klasifikasi optimal, diketahui bahwa PC2zg2 merupakan variabel pemilah yang utama dan paling penting dalam menentukan klasifikasi curah hujan di stasiun pengamatan Kemayoran. Pada Tabel 4.18, skor variabel PC2zg2 sebesar 100 karena mampu memberikan nilai penurunan keheterogenan tertinggi pada *node* utama. Selain itu, ada 31 variabel lain yang juga berkontribusi dalam pembentukan pohon klasifikasi optimal, hasil selengkapnya disajikan dalam Lampiran 13.

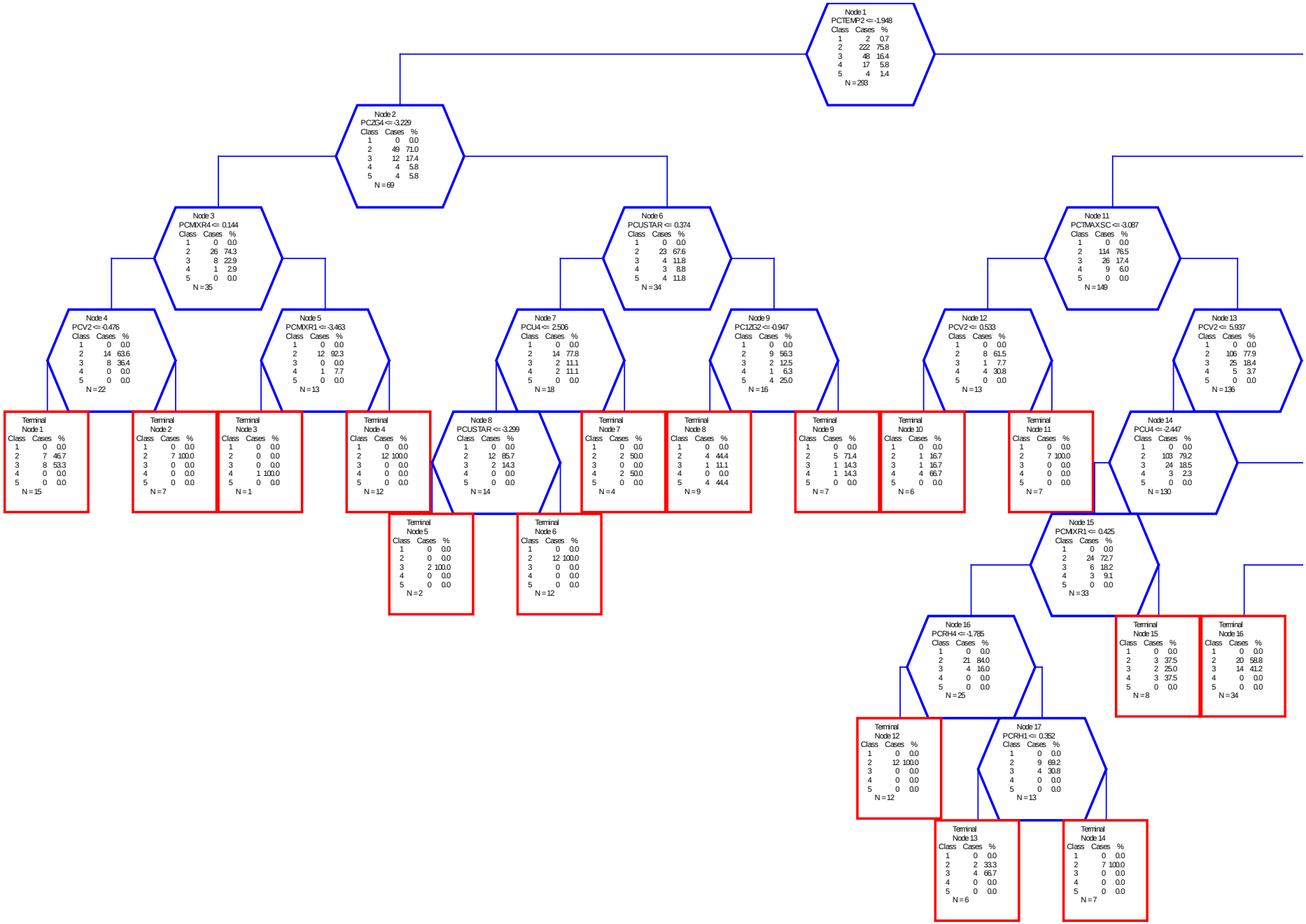
Tabel 4.18 Variabel Penting Pembentukan Pohon Klasifikasi Optimal Stasiun Kemayoran Sebelum SMOTE

<table>
<tr><th>Variabel</th><th>Skor Variabel</th><th></th></tr>
<tr><td>PC2zg2</td><td>100</td><td>||||||||||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCtemp2</td><td>95,57</td><td>||||||||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCrh1</td><td>93,98</td><td>|||||||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCmixr4</td><td>92,38</td><td>|||||||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCzg4</td><td>88,81</td><td>|||||||||||||||||||||||||||||||||||||</td></tr>
</table>

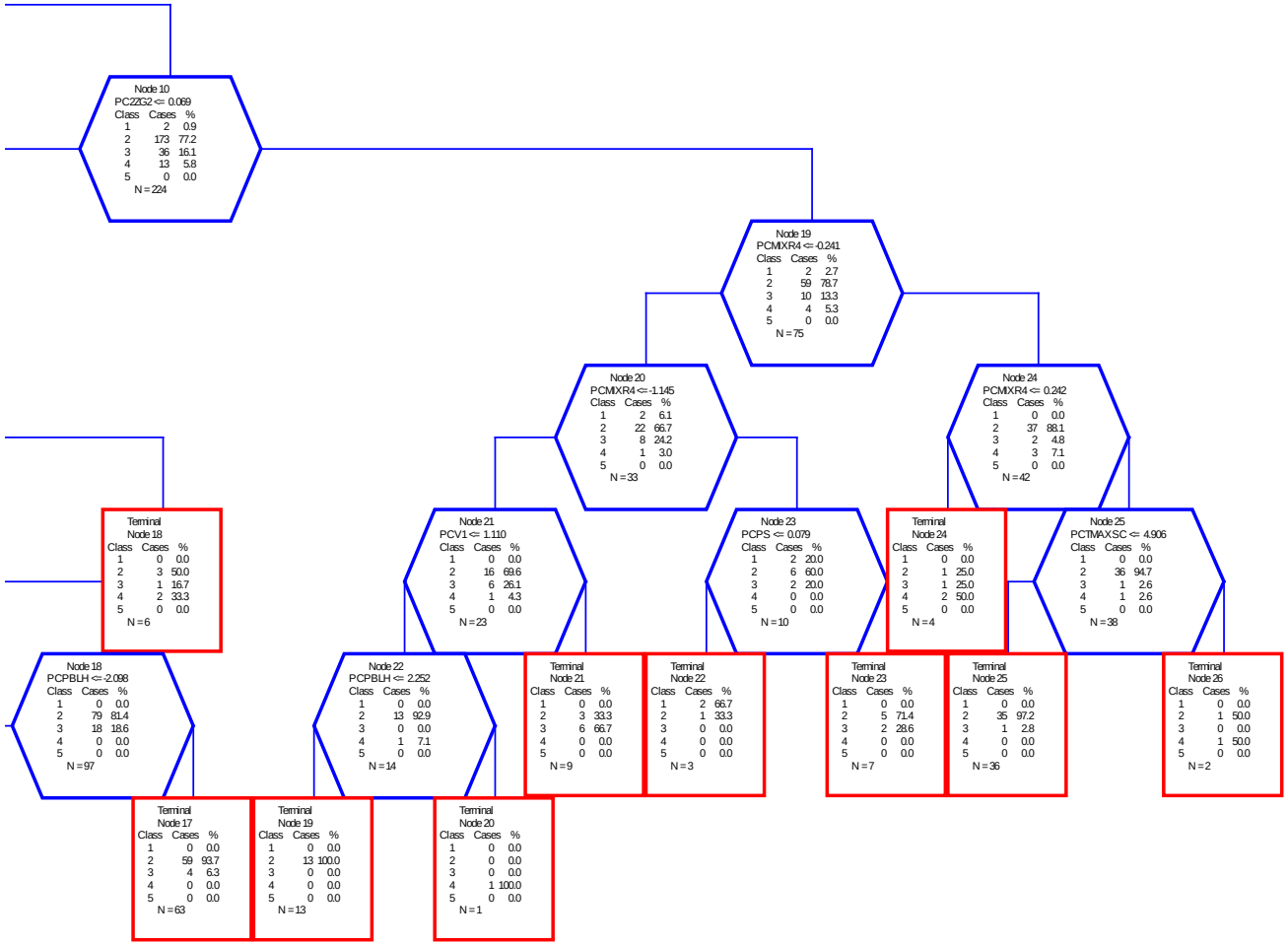
Walaupun variabel utama PC2zg2 mempunyai skor 100, namun pada struktur pohon optimal ditunjukkan bahwa variabel yang menjadi pemilah *node* 1 adalah variabel PCtemp2. Dimana jika nilai PCtemp2 $\leq$ -1,948 akan dipilah menjadi *node* kiri. Sedangkan jika nilai PCtemp2 > -1,948 akan dipilah menjadi *node* kanan. Gambar 4.9 merupakan visualisasi struktur pohon klasifikasi optimal.

Suatu *node* akan terus dipilah menjadi *node* anak baru (kiri dan kanan) sesuai prosedur *binary recursive partitioning*, sampai *node* tersebut telah dianggap memiliki anggota yang homogen atau jika *node* tersebut hanya memiliki 1 anggota pengamatan maka *node* akan menjadi *node* terminal dan tidak akan dipilah lagi. Pohon klasifikasi optimal yang terbentuk terdiri atas 26 *node* terminal seperti pada Gambar 4.9. Masing-masing *node* terminal tersebut memiliki karakteristik tertentu dan diprediksi sebagai kelas variabel respon tertentu sesuai dengan label kelas yang diberikan. Berdasarkan hasil penelusuran 26 terminal *node* tersebut, Tabel 4.19 menampilkan rangkuman pengklasifikasian curah hujan menurut indikasi kesamaan label kelas setiap *node* terminal.

(Halaman ini sengaja dikosongkan)



Gambar 4.9 Split Plot Pohon Optimal Stasiun Kemayoran Sebelum SMOTE


Node 10
PC2ZG2 <= 0.069
Class Cases %
1 2 0.9
2 173 77.2
3 36 16.1
4 13 5.8
5 0 0.0
N = 224
Node 19
PCMXR4 <= -0.241
Class Cases %
1 2 2.7
2 59 78.7
3 10 13.3
4 4 5.3
5 0 0.0
N = 75
Node 20
PCMXR4 <= -1.145
Class Cases %
1 2 6.1
2 22 66.7
3 8 24.2
4 1 3.0
5 0 0.0
N = 33
Node 24
PCMXR4 <= 0.242
Class Cases %
1 0 0.0
2 37 88.1
3 2 4.8
4 3 7.1
5 0 0.0
N = 42
Terminal
Node 18
Class Cases %
1 0 0.0
2 3 50.0
3 1 16.7
4 2 33.3
5 0 0.0
N = 6
Node 21
PCV1 <= 1.110
Class Cases %
1 0 0.0
2 16 69.6
3 6 26.1
4 1 4.3
5 0 0.0
N = 23
Node 23
PCPS <= 0.079
Class Cases %
1 2 20.0
2 6 60.0
3 2 20.0
4 0 0.0
5 0 0.0
N = 10
Terminal
Node 24
Class Cases %
1 0 0.0
2 1 25.0
3 1 25.0
4 2 50.0
5 0 0.0
N = 4
Node 25
PCTMAXSC <= 4.906
Class Cases %
1 0 0.0
2 36 94.7
3 1 2.6
4 1 2.6
5 0 0.0
N = 38
Node 18
PCPBLH <= -2.098
Class Cases %
1 0 0.0
2 79 81.4
3 18 18.6
4 0 0.0
5 0 0.0
N = 97
Node 22
PCPBLH <= 2.252
Class Cases %
1 0 0.0
2 13 92.9
3 0 0.0
4 1 7.1
5 0 0.0
N = 14
Terminal
Node 21
Class Cases %
1 0 0.0
2 3 33.3
3 6 66.7
4 0 0.0
5 0 0.0
N = 9
Terminal
Node 22
Class Cases %
1 2 66.7
2 1 33.3
3 0 0.0
4 0 0.0
5 0 0.0
N = 3
Terminal
Node 23
Class Cases %
1 0 0.0
2 5 71.4
3 2 28.6
4 0 0.0
5 0 0.0
N = 7
Terminal
Node 25
Class Cases %
1 0 0.0
2 35 97.2
3 1 2.8
4 0 0.0
5 0 0.0
N = 36
Terminal
Node 26
Class Cases %
1 0 0.0
2 1 50.0
3 0 0.0
4 1 50.0
5 0 0.0
N = 2
Terminal
Node 17
Class Cases %
1 0 0.0
2 59 93.7
3 4 6.3
4 0 0.0
5 0 0.0
N = 63
Terminal
Node 19
Class Cases %
1 0 0.0
2 13 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 13
Terminal
Node 20
Class Cases %
1 0 0.0
2 0 0.0
3 0 0.0
4 1 100.0
5 0 0.0
N = 1

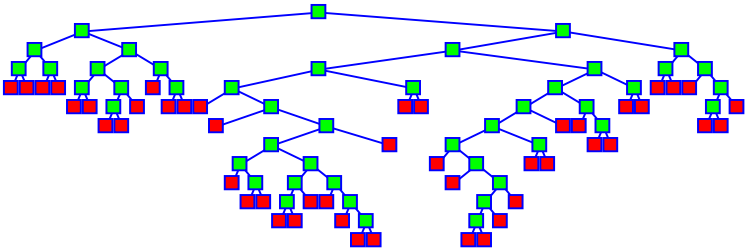
Tabel 4.19 Kelas Curah Hujan Stasiun Kemayoran pada Masing-Masing Terminal *Node* Sebelum SMOTE

Kelas	*Terminal Node*	Persentase	*Terminal Node*	Persentase
1	33	100		
2	2	100	14	100
	4	100	17	76,1
	6	100	19	100
	11	100	25	88,3
	12	100		
3	1	84,1	16	76,4
	5	100	21	90,2
	13	90,2	23	64,9
4	3	100	18	77,4
	7	92,9	20	100
	9	57,6	24	82,3
	10	90,3	26	92,9
	15	76,2		
5	4	96,3		

Dari Tabel 4.19 dapat lihat bahwa kelas 2 dan kelas 4 merupakan kelas klasifikasi yang paling banyak terbentuk. Hal ini disebabkan distribusi data curah hujan tidak *balance* antara cerah berawan, hujan ringan, hujan sedang, hujan lebat, dan hujan lebat sekali. Pada pengamatan stasiun Kemayoran data didominasi dengan hujan ringan (kelas 2) dan jumlah data paling rendah berada pada kejadian cerah berawan (kelas 1) dan hujan lebat sekali (kelas 5). Sehingga pada Tabel 4.19, kelas 1 dan kelas 5 hanya terdapat pada 1 terminal *node*. Maka dari itu, pada stasiun Kemayoran akan dicobakan proses SMOTE untuk mengatasi masalah data yang tidak *balance*.

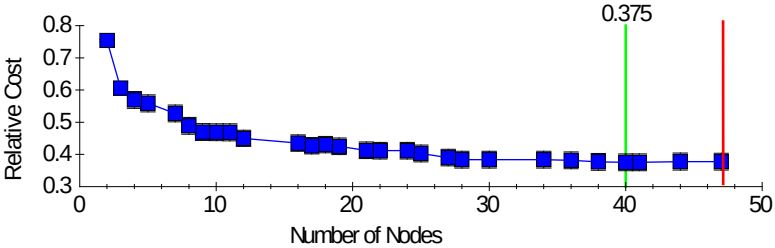
Proses SMOTE pada stasiun Kemayoran dilakukan sebanyak 12 iterasi dan menambah jumlah data pengamatan

menjadi 466 pengamatan. Kemudian data pengamatan tersebut digunakan untuk membangun model klasifikasi pohon yang baru. Pohon mkasimal yang dihasilkan memiliki 47 terminal *node* dengan kedalaman sebesar 12 tingkatan. Variabel yang menjadi pemilah utama adalah variabel PCtemp2 dengan skor 100. Topologi pohon maksimal ditampilkan pada Gambar 4.10.



Gambar 4.10 Topologi Pohon Maksimal Stasiun Kemayoran Setelah SMOTE

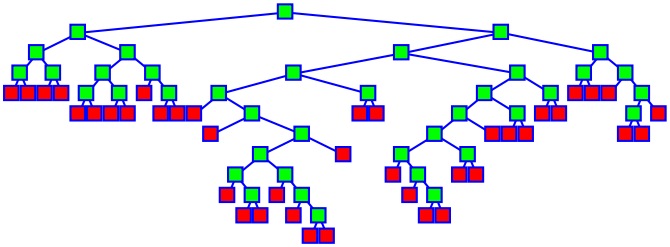
Gambar 4.11 menampilkan adanya perbedaan nilai *relative cost* yang dihasilkan oleh pohon klasifikasi maksimal dengan pohon klasifikasi yang dianggap optimal. Pohon klasifikasi maksimal ditunjukkan oleh garis berwarna merah sedangkan pohon klasifikasi optimal ditunjukkan oleh garis berwarna hijau.



Gambar 4.11 Plot *Relative Cost* Klasifikasi Curah Hujan Stasiun Kemayoran Setelah SMOTE

Berdasarkan Gambar 4.11, pohon klasifikasi maksimal yang terbentuk terdiri dari 47 terminal *nodes* dan *relative cost* sebesar 0,377 ± 0,026. Pemangkasan pohon dilakukan secara iteratif berdasarkan *cross validated relative cost* yang minimum. Nilai *cross validated relative cost* yang minimum adalah 0,375 ±

0,026 pada saat *terminal nodes* sebanyak 40. Sehingga dapat dikatakan bahwa pohon klasifikasi optimal yang terbentuk terdiri dari 40 *terminal nodes*. Karena nilai *relative cost* pohon klasifikasi optimal lebih kecil maka pohon klasifikasi optimal dipilih sebagai pohon yang layak untuk pohon klasifikasi curah hujan pada stasiun pengamatan Kemayoran.



Gambar 4.12 Topologi Pohon Optimal Stasiun Kemayoran Setelah SMOTE

Berdasarkan topologi pohon klasifikasi optimal, diketahui bahwa PCtemp2 merupakan variabel pemilah yang utama dan paling penting dalam menentukan klasifikasi curah hujan di stasiun pengamatan Kemayoran. Pada Tabel 4.20, skor variabel PCtemp2 adalah 100 karena mampu memberikan nilai penurunan keheterogenan tertinggi pada *node* utama. Selain itu ada 31 variabel lain yang juga berkontribusi dalam pembentukan pohon klasifikasi optimal, hasil selengkapnya disajikan dalam Lampiran 15.

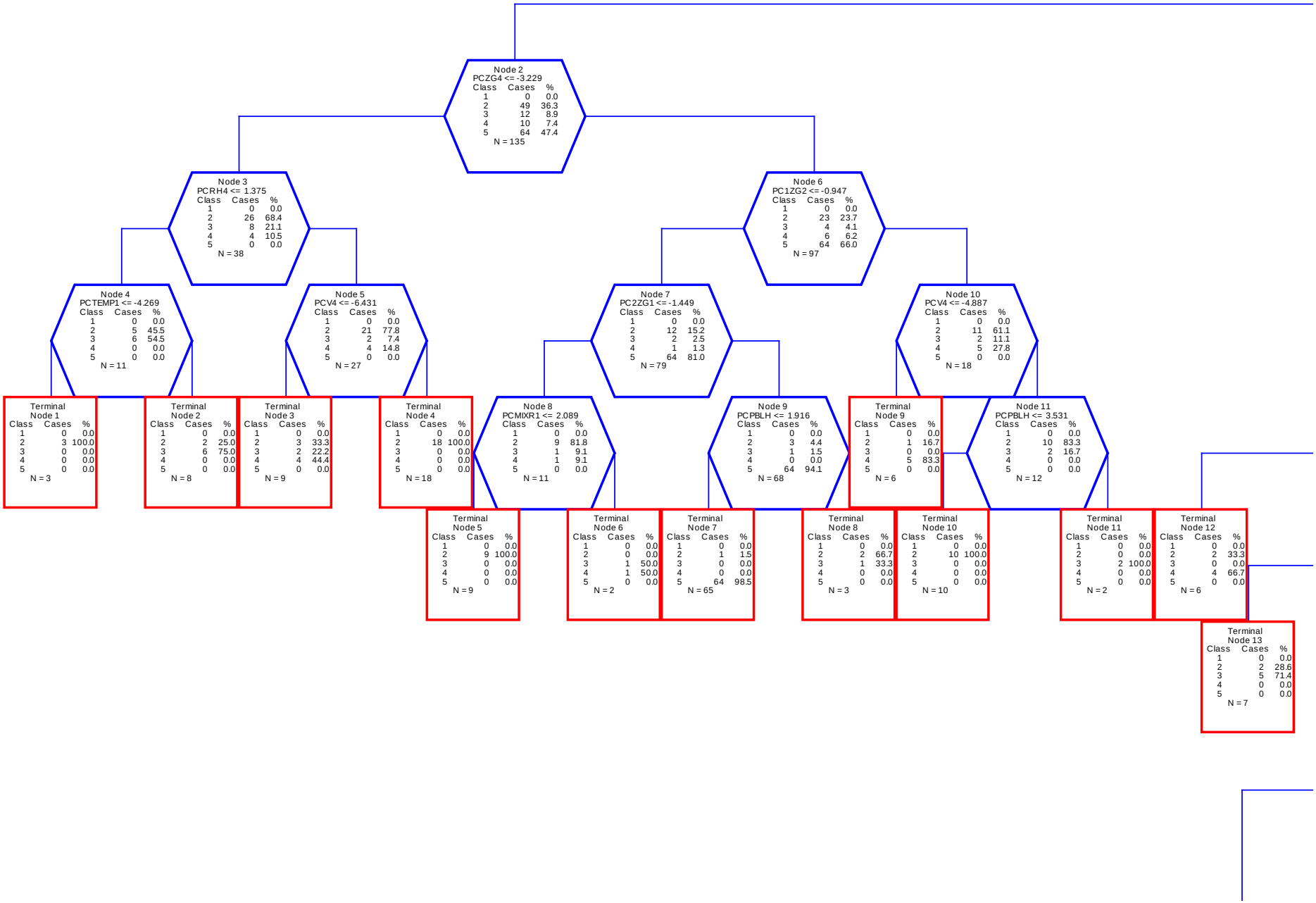
Tabel 4.20 Variabel Penting Pembentukan Pohon Klasifikasi Optimal Stasiun Kemayoran Setelah SMOTE

Variabel	**Skor Variabel**	
PCtemp2	100	||||||||||||||||||||||||||||||||||||||||||
PCzg4	95,43	||||||||||||||||||||||||||||||||||||||||
PC1zg2	85,46	||||||||||||||||||||||||||||||||||||
PCu4	78,88	|||||||||||||||||||||||||||||||||
PCrnd	71,88	||||||||||||||||||||||||||||||

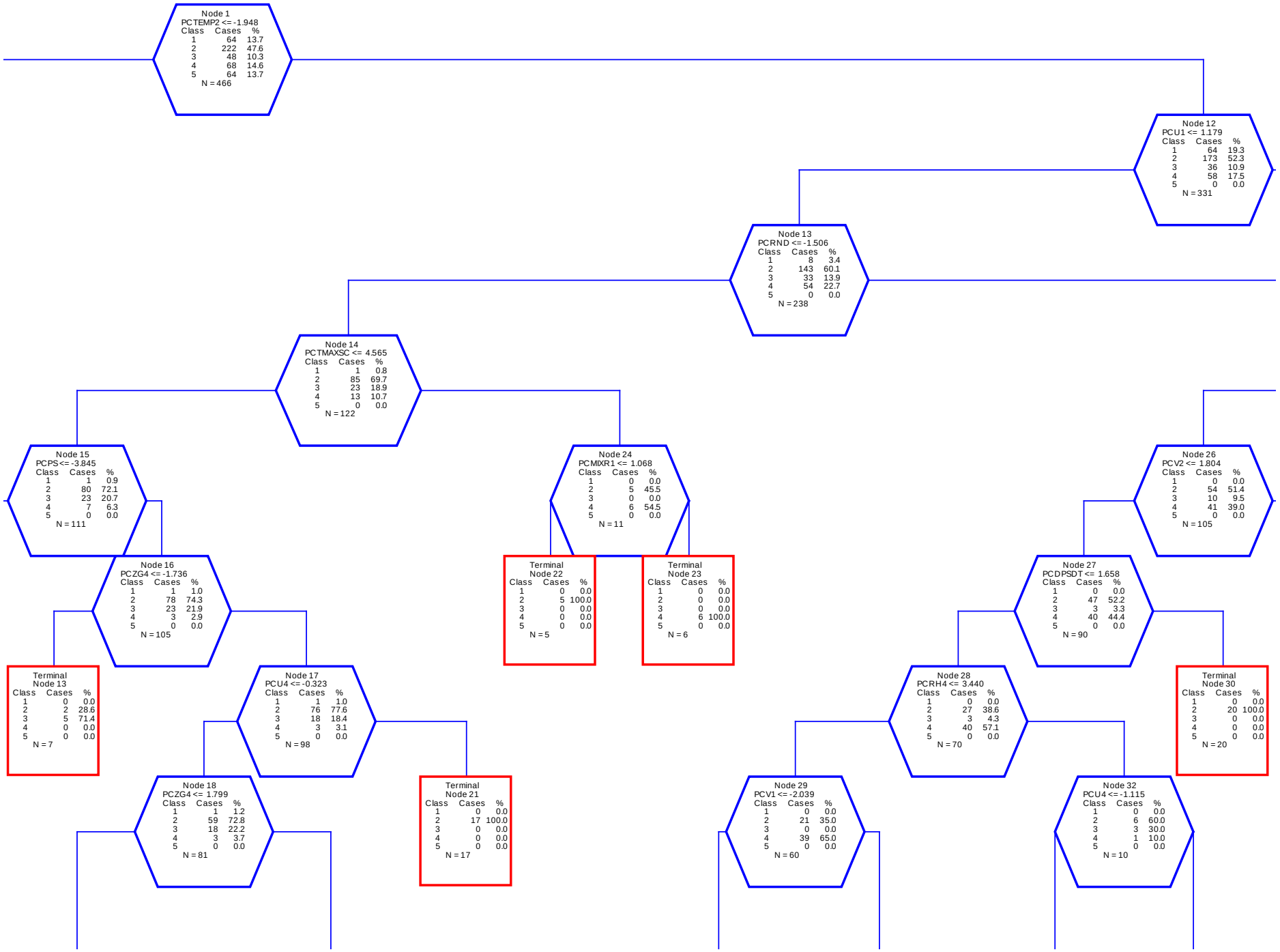
Variabel utama PCtemp2 memilah *node* utama menjadi *node* kanan dan kiri dengan ketentuan nilai PCtemp2 $\leq$ -1,948 akan dipilah menjadi *node* kiri. Sedangkan jika nilai PCtemp2 > -

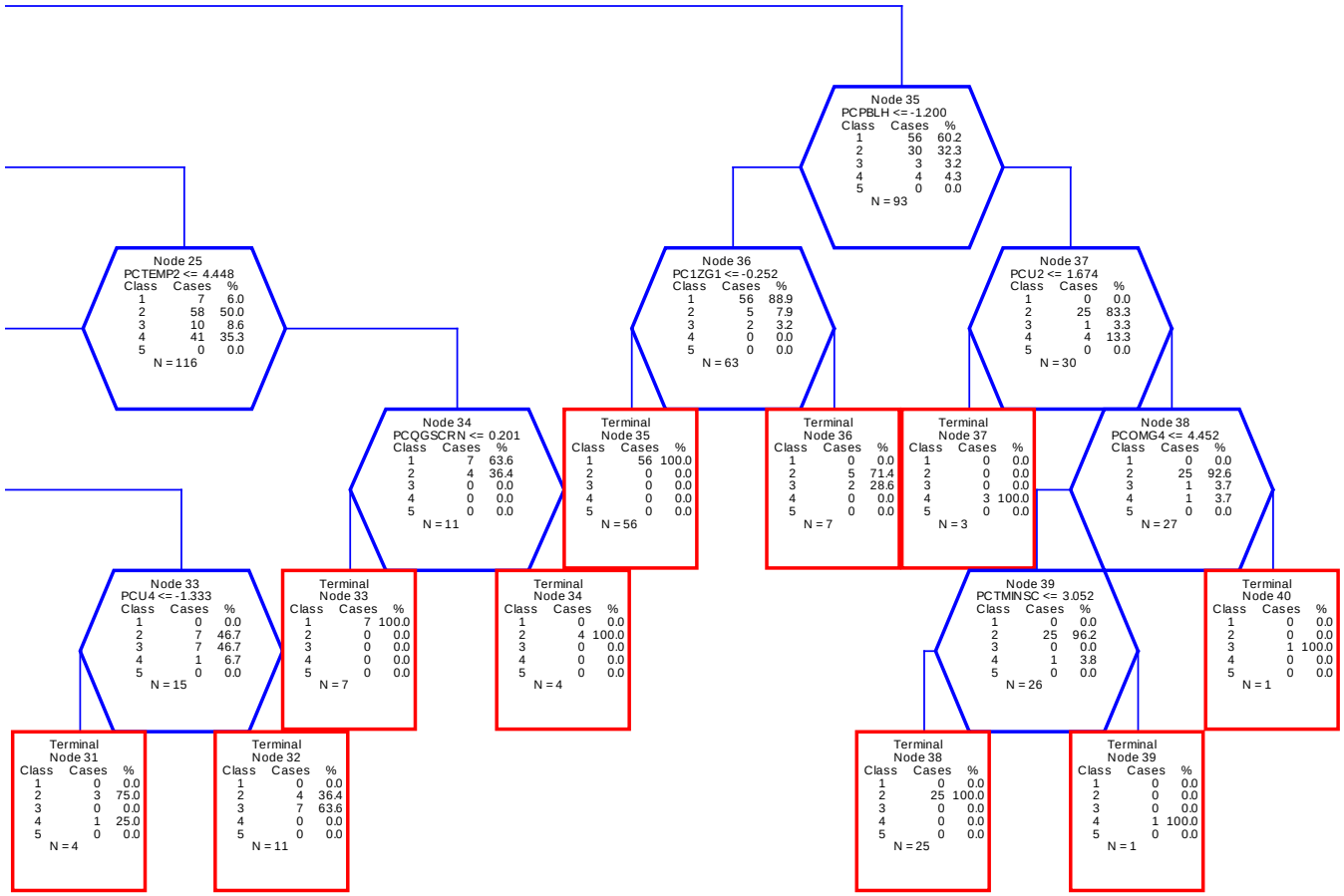
1,948 akan dipilah menjadi *node* kanan. Gambar 4.13 merupakan visualisasi struktur pohon klasifikasi optimal.

Suatu *node* akan terus dipilah menjadi *node* anak baru (kiri dan kanan) sesuai prosedur *binary recursive partitioning*, sampai *node* tersebut telah dianggap memiliki anggota yang homogen atau jika *node* tersebut hanya memiliki 1 anggota pengamatan maka *node* akan menjadi *node* terminal dan tidak akan dipilah lagi. Pohon klasifikasi optimal yang terbentuk terdiri atas 40 *node* terminal seperti pada Gambar 4.13. Masing-masing *node* terminal tersebut memiliki karakteristik tertentu dan diprediksi sebagai kelas variabel respon tertentu sesuai dengan label kelas yang diberikan. Berdasarkan hasil penelusuran 40 terminal *node* tersebut, Tabel 4.21 menampilkan rangkuman pengklasifikasian curah hujan menurut indikasi kesamaan label kelas setiap *node* terminal.



Gambar 4.13 Split Plot Pohon Optimal Stasiun Kemayoran Setelah SMOTE


Node 1
PCTEMP2 <= -1.948
Class Cases %
1 64 13.7
2 222 47.6
3 48 10.3
4 68 14.6
5 64 13.7
N = 466
Node 12
PCU1 <= 1.179
Class Cases %
1 64 19.3
2 173 52.3
3 36 10.9
4 58 17.5
5 0 0.0
N = 331
Node 13
PCRND <= -1.506
Class Cases %
1 8 3.4
2 143 60.1
3 33 13.9
4 54 22.7
5 0 0.0
N = 238
Node 14
PCTMAXSC <= 4.565
Class Cases %
1 1 0.8
2 85 69.7
3 23 18.9
4 13 10.7
5 0 0.0
N = 122
Node 15
PCPS <= -3.845
Class Cases %
1 1 0.9
2 80 72.1
3 23 20.7
4 7 6.3
5 0 0.0
N = 111
Node 24
PCMIXR1 <= 1.068
Class Cases %
1 0 0.0
2 5 45.5
3 0 0.0
4 6 54.5
5 0 0.0
N = 11
Node 26
PCV2 <= 1.804
Class Cases %
1 0 0.0
2 54 51.4
3 10 9.5
4 41 39.0
5 0 0.0
N = 105
Node 16
PCZG4 <= -1.736
Class Cases %
1 1 1.0
2 78 74.3
3 23 21.9
4 3 2.9
5 0 0.0
N = 105
Terminal
Node 22
Class Cases %
1 0 0.0
2 5 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 5
Terminal
Node 23
Class Cases %
1 0 0.0
2 0 0.0
3 0 0.0
4 6 100.0
5 0 0.0
N = 6
Node 27
PCDPSDT <= 1.658
Class Cases %
1 0 0.0
2 47 52.2
3 3 3.3
4 40 44.4
5 0 0.0
N = 90
Terminal
Node 13
Class Cases %
1 0 0.0
2 2 28.6
3 5 71.4
4 0 0.0
5 0 0.0
N = 7
Node 17
PCU4 <= -0.323
Class Cases %
1 1 1.0
2 76 77.6
3 18 18.4
4 3 3.1
5 0 0.0
N = 98
Node 28
PCRH4 <= 3.440
Class Cases %
1 0 0.0
2 27 38.6
3 3 4.3
4 40 57.1
5 0 0.0
N = 70
Terminal
Node 30
Class Cases %
1 0 0.0
2 20 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 20
Node 18
PCZG4 <= 1.799
Class Cases %
1 1 1.2
2 59 72.8
3 18 22.2
4 3 3.7
5 0 0.0
N = 81
Terminal
Node 21
Class Cases %
1 0 0.0
2 17 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 17
Node 29
PCV1 <= -2.039
Class Cases %
1 0 0.0
2 21 35.0
3 0 0.0
4 39 65.0
5 0 0.0
N = 60
Node 32
PCU4 <= -1.115
Class Cases %
1 0 0.0
2 6 60.0
3 3 30.0
4 1 10.0
5 0 0.0
N = 10


Node 35
PCPBLH <= -1.200
Class Cases %
1 56 60.2
2 30 32.3
3 3 3.2
4 4 4.3
5 0 0.0
N = 93
Node 25
PCTEMP2 <= 4.448
Class Cases %
1 7 6.0
2 58 50.0
3 10 8.6
4 41 35.3
5 0 0.0
N = 116
Node 36
PC1ZG1 <= -0.252
Class Cases %
1 56 88.9
2 5 7.9
3 2 3.2
4 0 0.0
5 0 0.0
N = 63
Node 37
PCU2 <= 1.674
Class Cases %
1 0 0.0
2 25 83.3
3 1 3.3
4 4 13.3
5 0 0.0
N = 30
Node 34
PCQGSCRN <= 0.201
Class Cases %
1 7 63.6
2 4 36.4
3 0 0.0
4 0 0.0
5 0 0.0
N = 11
Terminal Node 35
Class Cases %
1 56 100.0
2 0 0.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 56
Terminal Node 36
Class Cases %
1 0 0.0
2 5 71.4
3 2 28.6
4 0 0.0
5 0 0.0
N = 7
Terminal Node 37
Class Cases %
1 0 0.0
2 0 0.0
3 0 0.0
4 3 100.0
5 0 0.0
N = 3
Node 38
PCOMG4 <= 4.452
Class Cases %
1 0 0.0
2 25 92.6
3 1 3.7
4 1 3.7
5 0 0.0
N = 27
Node 33
PCU4 <= -1.333
Class Cases %
1 0 0.0
2 7 46.7
3 7 46.7
4 1 6.7
5 0 0.0
N = 15
Terminal Node 33
Class Cases %
1 7 100.0
2 0 0.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 7
Terminal Node 34
Class Cases %
1 0 0.0
2 4 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 4
Node 39
PCTMINSC <= 3.052
Class Cases %
1 0 0.0
2 25 96.2
3 0 0.0
4 1 3.8
5 0 0.0
N = 26
Terminal Node 40
Class Cases %
1 0 0.0
2 0 0.0
3 1 100.0
4 0 0.0
5 0 0.0
N = 1
Terminal Node 31
Class Cases %
1 0 0.0
2 3 75.0
3 0 0.0
4 1 25.0
5 0 0.0
N = 4
Terminal Node 32
Class Cases %
1 0 0.0
2 4 36.4
3 7 63.6
4 0 0.0
5 0 0.0
N = 11
Terminal Node 38
Class Cases %
1 0 0.0
2 25 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 25
Terminal Node 39
Class Cases %
1 0 0.0
2 0 0.0
3 0 0.0
4 1 100.0
5 0 0.0
N = 1

Terminal
Node 14

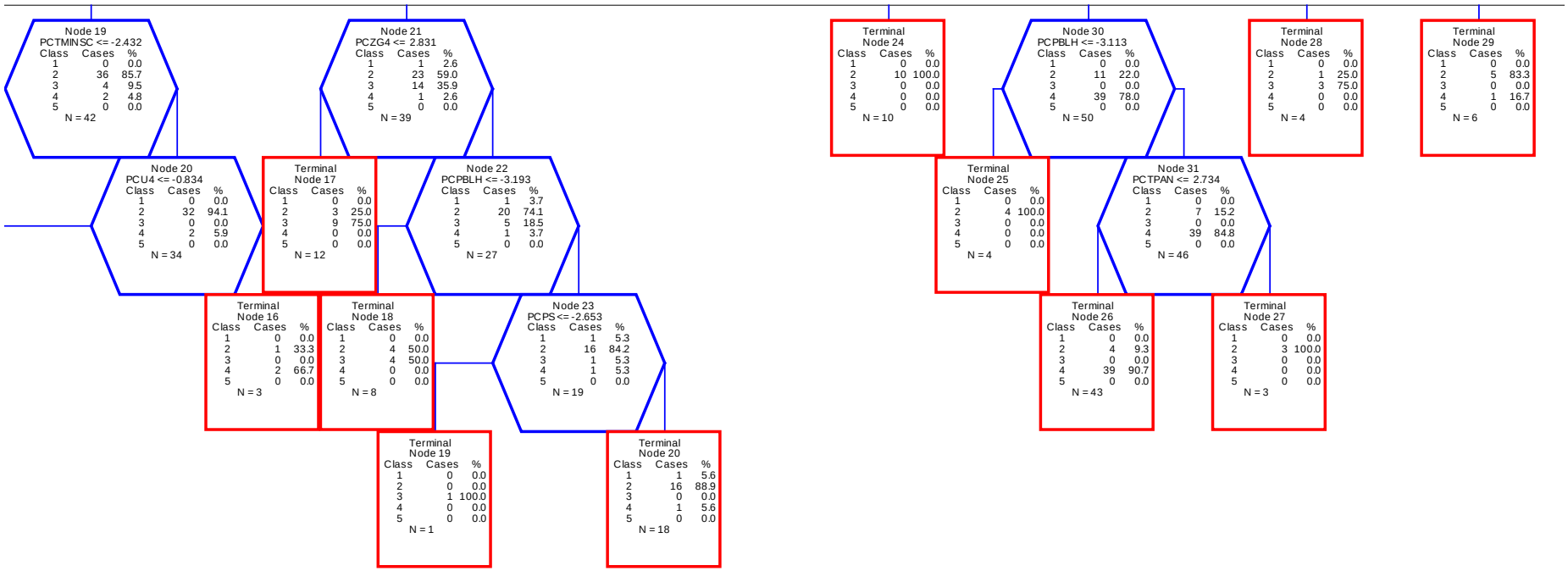
Class	Cases	%
1	0	0.0
2	4	50.0
3	4	50.0
4	0	0.0
5	0	0.0

N = 8

Terminal
Node 15

Class	Cases	%
1	0	0.0
2	31	100.0
3	0	0.0
4	0	0.0
5	0	0.0

N = 31


Node 19
PCTMINSC <= -2.432
Class Cases %
1 0 0.0
2 36 85.7
3 4 9.5
4 2 4.8
5 0 0.0
N = 42
Node 21
PCZG4 <= 2.831
Class Cases %
1 1 2.6
2 23 59.0
3 14 35.9
4 1 2.6
5 0 0.0
N = 39
Terminal
Node 24
Class Cases %
1 0 0.0
2 10 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 10
Node 30
PCPBLH <= -3.113
Class Cases %
1 0 0.0
2 11 22.0
3 0 0.0
4 39 78.0
5 0 0.0
N = 50
Terminal
Node 28
Class Cases %
1 0 0.0
2 1 25.0
3 3 75.0
4 0 0.0
5 0 0.0
N = 4
Terminal
Node 29
Class Cases %
1 0 0.0
2 5 83.3
3 0 0.0
4 1 16.7
5 0 0.0
N = 6
Node 20
PCU4 <= -0.834
Class Cases %
1 0 0.0
2 32 94.1
3 0 0.0
4 2 5.9
5 0 0.0
N = 34
Terminal
Node 17
Class Cases %
1 0 0.0
2 3 25.0
3 9 75.0
4 0 0.0
5 0 0.0
N = 12
Node 22
PCPBLH <= -3.193
Class Cases %
1 1 3.7
2 20 74.1
3 5 18.5
4 1 3.7
5 0 0.0
N = 27
Terminal
Node 25
Class Cases %
1 0 0.0
2 4 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 4
Node 31
PCTPAN <= 2.734
Class Cases %
1 0 0.0
2 7 15.2
3 0 0.0
4 39 84.8
5 0 0.0
N = 46
Terminal
Node 16
Class Cases %
1 0 0.0
2 1 33.3
3 0 0.0
4 2 66.7
5 0 0.0
N = 3
Terminal
Node 18
Class Cases %
1 0 0.0
2 4 50.0
3 4 50.0
4 0 0.0
5 0 0.0
N = 8
Node 23
PCPS <= -2.653
Class Cases %
1 1 5.3
2 16 84.2
3 1 5.3
4 1 5.3
5 0 0.0
N = 19
Terminal
Node 26
Class Cases %
1 0 0.0
2 4 9.3
3 0 0.0
4 39 90.7
5 0 0.0
N = 43
Terminal
Node 27
Class Cases %
1 0 0.0
2 3 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 3
Terminal
Node 19
Class Cases %
1 0 0.0
2 0 0.0
3 1 100.0
4 0 0.0
5 0 0.0
N = 1
Terminal
Node 20
Class Cases %
1 1 5.6
2 16 88.9
3 0 0.0
4 1 5.6
5 0 0.0
N = 18

Tabel 4.21 Kelas Curah Hujan Stasiun Kemayoran pada Masing-Masing Terminal *Node* Setelah SMOTE

Kelas	*Terminal Node*	Persentase	*Terminal Node*	Persentase
1	33	100	35	100
2	1	100	24	100
	4	100	25	100
	5	100	27	100
	10	100	29	61
	15	100	30	100
	20	70	34	100
	21	100	38	100
	22	100		
3	2	93	18	82
	6	59	19	100
	8	70	28	93
	11	100	32	89
	13	92	36	65
	14	82	40	100
	17	93		
4	3	52	26	97
	9	94	31	52
	12	87	37	100
	16	87	39	100
	23	100		
5	7	100		

Dari Tabel 4.21 dapat diketahui walaupun sudah dilakukan proses SMOTE untuk mengatasi data *imbalance*, tetapi terminal *node* yang terbentuk masih cenderung pada kelas 2. Karena pada setelah dilakukan proses SMOTE, data pada kelas 2 masih tetap paling tinggi dibandingkan kelas lainnya. Secara keseluruhan

dapat diketahui bahwa terdapat 63 pengamatan dalam kelas cerah berawan, 174 pengamatan masuk dalam kelas hujan ringan, 55 pengamatan termasuk dalam kelas hujan sedang, 68 pengamatan masuk dalam kelas hujan lebat dan 64 pengamatan yang termasuk dalam kelas hujan lebat sekali.

Penelusuran struktur pohon klasifikasi optimal terhadap *node* terminal dapat memberikan informasi tentang karakteristik kelas *node* terminal dengan persentase tertinggi untuk masing-masing kelas. Karakteristik kelas curah hujan pada masing-masing *node* terminal disajikan pada Tabel 4.22.

Tabel 4.22 Karakteristik Kelas Curah Hujan Stasiun Kemayoran Setelah SMOTE

Kelas	**Karakteristik**
Cerah Berawan (1)	PCtemp2 > -1,948 PCu1 ≤ 1,179 PCrnd > -1,505 PCtemp2 > 4,448 PCqgscrn ≤ 0,201
Hujan Ringan (2)	PCtemp2 ≤ -1,948 PCzg4 ≤ -3,228 PCrh4 ≤ 1,375 PCtemp1 ≤ -4,268
Hujan Sedang (3)	PCtemp2 ≤ -1,948 PCzg4 > -3,228 PC1zg2 > -0,946 PCv4 > -4,886 PCpblh >3,531
Hujan Lebat (4)	PCtemp2 > -1,948 PCu1 ≤ 1,179 PCrnd ≤ -1,505 PCtmaxscr > 4,564 PCmixr1 > 1,067
Hujan Lebat Sekali (5)	PCtemp2 ≤ -1,948 PCzg4 > -3,228 PC1zg2 ≤ -0,946 PC2zg1 > -1,448 PCpblh ≤ 1,915

4.3.2.4 Hasil Ketepatan Klasifikasi Klasifikasi Pohon

Tingkat keakuratan hasil klasifikasi pohon optimal yang dihasilkan dari data *learning* dapat dihitung berdasarkan Tabel 4.23.

Tabel 4.23 Klasifikasi Curah Hujan Data *Learning* pada Pohon Optimal Stasiun Kemayoran Sebelum SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	0	1	1	0	0	0	2
2	3	120	63	31	5	54,05	102
3	2	16	18	11	1	37,50	30
4	1	6	6	3	1	17,65	14
5	0	3	1	0	0	0	4

Berdasarkan Tabel 4.23, kesalahan klasifikasi kelas pengamatan terjadi pada seluruh kelas. Pada kelas 1 dan kelas 5 tidak ada 1 pun pengamatan yang diklasifikasikan dengan tepat. Sebanyak 102 pengamatan yang secara aktual termasuk kelas 2 (hujan ringan) namun salah diklasifikasikan sebagai sebagai kelas 1 (cerah berawan), 3 (hujan sedang), 4 (hujan lebat) dan 5 (hujan lebat sekali). Kesalahan klasifikasi juga terjadi pada kelas 3 (hujan sedang) dimana sebanyak 30 pengamatan berada pada kelas 1, 2, 4 dan 5. Selanjutnya sebanyak 14 pengamatan yang secara aktual masuk kelas 4 (hujan lebat), namun salah diklasifikasikan sebagai kelas 1 (cerah berawan), 2 (hujan ringan), 3 (hujan sedang), dan 5 (hujan lebat sekali).

Menggunakan informasi pada Tabel 4.23, maka ketepatan klasifikasi data *learning* dapat dihitung sebagai berikut:

$$1 - APER = \left(1 - \frac{2 + 102 + 30 + 14 + 4}{293}\right) \times 100\% = 48{,}12\%$$

Hasil perhitungan ketepatan klasifikasi data *learning* sebesar 48,12 persen. Artinya pohon klasifikasi optimal mampu mengklasifikasikan pengamatan curah hujan kedalam kelas kategori hujan dengan tepat sebesar 48,12 persen.

Pohon klasifikasi optimal yang terbentuk perlu divalidasi untuk mengetahui apakah pohon klasifikasi tersebut layak dan dapat digunakan untuk mengklasifikasi data baru. Tingkat keakuratan hasil klasifikasi pohon optimal yang dihasilkan dari data *testing* dapat dihitung berdasarkan Tabel 4.24.

Tabel 4.24 Klasifikasi Curah Hujan Data *Testing* pada Pohon Optimal Stasiun Kemayoran Sebelum SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	0	0	0	0	0	0 %	0
2	0	6	0	1	0	85,71 %	1
3	0	0	0	0	0	0 %	0
4	0	0	0	0	0	0 %	0
5	0	0	0	0	0	0 %	0

Sehingga dapat dihitung besarnya ketepatan klasifikasi untuk data *testing* sebagai berikut:

$$1 - APER = \left(1 - \frac{6}{7}\right) \times 100\% = 85{,}71\%$$

Sedangkan untuk ketepatan klasifikasi setelah dilakukan proses SMOTE ditampilkan pada Tabel 4.25.

Tabel 4.25 Klasifikasi Curah Hujan Data *Learning* pada Pohon Optimal Stasiun Kemayoran Setelah SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	60	2	1	1	0	94	4
2	1	124	61	30	6	56	98
3	1	19	18	9	1	38	30
4	2	12	6	47	1	69	21
5	0	1	1	2	60	9	4

Berdasarkan Tabel 4.25, kesalahan klasifikasi kelas pengamatan terjadi pada seluruh kelas. Pada kelas 1 dan kelas 5 terdapat 4 pengamatan yang salah dikalsifikan. Sebanyak 98 pengamatan yang secara aktual termasuk kelas 2 (hujan ringan) namun salah diklasifikasikan sebagai sebagai kelas 1 (cerah berawan), 3 (hujan sedang), 4 (hujan lebat) dan 5 (hujan lebat

sekali). Kesalahan klasifikasi juga terjadi pada kelas 3 (hujan sedang) dimana sebanyak 30 pengamatan berada pada kelas 1, 2, 4 dan 5. Selanjutnya sebanyak 21 pengamatan yang secara aktual masuk kelas 4 (hujan lebat), namun salah diklasifikasikan sebagai kelas 1 (cerah berawan), 2 (hujan ringan), 3 (hujan sedang), dan 5 (hujan lebat sekali).

Menggunakan informasi pada Tabel 4.25, maka ketepatan klasifikasi data *learning* dapat dihitung sebagai berikut:

$$1 - APER = \left(1 - \frac{4 + 98 + 30 + 21 + 4}{466}\right) \times 100\% = 66{,}3\%$$

Hasil perhitungan ketepatan klasifikasi data *learning* sebesar 66,3 persen. Artinya pohon klasifikasi optimal mampu mengklasifikasikan pengamatan curah hujan kedalam kelas kategori hujan dengan tepat sebesar 66,3 persen.

Pohon klasifikasi optimal yang terbentuk perlu divalidasi untuk mengetahui apakah pohon klasifikasi tersebut layak dan dapat digunakan untuk mengklasifikasi data baru. Tingkat keakuratan hasil klasifikasi pohon optimal yang dihasilkan dari data *testing* dapat dihitung berdasarkan Tabel 4.26.

Tabel 4.26 Klasifikasi Curah Hujan Data *Testing* pada Pohon Optimal Stasiun Kemayoran Setelah SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	0	0	0	0	0	0	0
2	0	6	0	1	0	85,71	1
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0

Sehingga dapat dihitung besarnya ketepatan klasifikasi untuk data *testing* setelah SMOTE sebagai berikut:

$$1 - APER = \left(1 - \frac{6}{7}\right) \times 100\% = 85.71\%$$

Berikut adalah perbandingan hasil ketepatan klasifikasi pohon maksimal dengan pohon optimal yang ditunjukkan oleh Tabel 4.27.

Tabel 4.27 Perbandingan Ketepatan Klasifikasi Pohon Maksimal dan Pohon Optimal Stasiun Kemayoran

	Pohon Klasifikasi	**Ketepatan Klasifikasi (%)**	
		Learning	***Testing* Data Baru**
Sebelum SMOTE	Pohon Maksimal	51,19	
	Pohon Optimal	48,12	85,71
Setelah SMOTE	Pohon Maksimal	67,20	
	Pohon Optimal	66,30	85,71

Berdasarkan Tabel 4.27, dapat diketahui bahwa ketepatan klasifikasi pohon maksimal lebih tinggi daripada pohon optimal. Hal ini dikarenakan pohon klasifikasi maksimal memiliki *node* yang paling banyak dengan melibatkan lebih banyak variabel prediktor sebagai pemilah *node* sehingga kemungkinan klasifikasi data dengan tepat akan cenderung lebih besar. Hasil ketepatan klasifikasi data *testing* pada pohon optimal setelah dan sebelum SMOTE memiliki nilai yang sama yakni 85,71 persen. Sedangkan untuk *cross validation* pohon optimal setelah SMOTE menunjukkan peningkatan dari 48,12% menjadi 66,3%. Artinya setelah dilakukan SMOTE, pohon optimal yang dihasilkan lebih baik daripada sebelum dilakukan SMOTE.

4.3.3 Klasifikasi Curah Hujan Stasiun Pondok Betung

Variabel prediktor yang digunakan sebanyak 37 komponen utama dengan jumlah data 384 pengamatan. Data *testing* yang digunakan adalah data *testing* yang ditetapkan pada saat proses reduksi dimensi dengan PCA. Analisis klasifikasi pohon diawali dengan pembentukkan pohon klasifikasi maksimal. Berikut penjelasan masing-masing tahapan analisis klasifikasi pohon

dengan menggunakan kombinasi data *learning* dan data *testing* tersebut.

4.3.3.1 Pembentukan Pohon Klasifikasi Maksimal

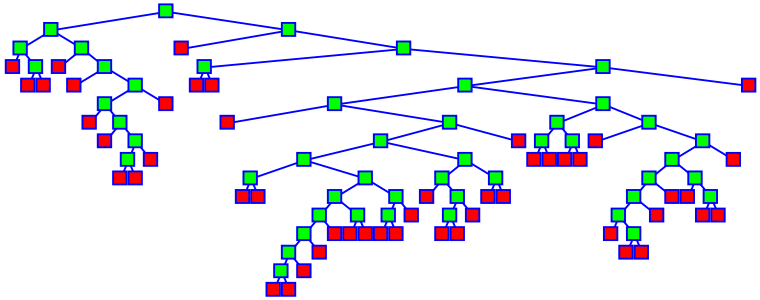
Tahap awal yang dilakukan untuk membentuk pohon klasifikasi adalah dengan menentukan variabel pemilah. Variabel pemilah dipilih dari beberapa kemungkinan pemilah dari masing-masing variabel. Pemilah yang terpilih adalah variabel pemilah dan nilai variabel (*threshold*) yang memiliki nilai *goodness of split* tertinggi. Pemilah yang terpilih merupakan variabel yang terpenting dalam mengklasifikasikan data pengamatan. Besarnya kontribusi variabel sebagai pemilah baik pemilah utama maupun pengganti pada pohon klasifikasi maksimal yang terbentuk ditunjukkan melalui suatu angka skor yang ditampilkan pada Tabel 4.28, selengkapnya pada Lampiran 18.

Tabel 4.28 Variabel Penting Pembentukan Pohon Klasifikasi Maksimal Stasiun Pondok Betung Sebelum SMOTE

Variabel	Skor Variabel	
PCustar	100	||||||||||||||||||||||||||||||||||||||||||
PCtscrn	78,05	|||||||||||||||||||||||||||||||||
Pctpan	76,86	||||||||||||||||||||||||||||||||
PCtemp1	72,44	||||||||||||||||||||||||||||||
PCtemp2	64,20	|||||||||||||||||||||||||||
PC1zg4	62,14	||||||||||||||||||||||||||
PCv1	56,96	|||||||||||||||||||||||
PCu1	51,82	|||||||||||||||||||||
PC1zg2	50,44	|||||||||||||||||||||

Tabel 4.28 menunjukkan bahwa berdasarkan skor yang dihasilkan diketahui variabel yang terpenting dan menjadi pemilah utama dalam mengklasifikasikan curah hujan adalah PCustar karena memiliki skor paling tinggi yaitu sebesar 100.

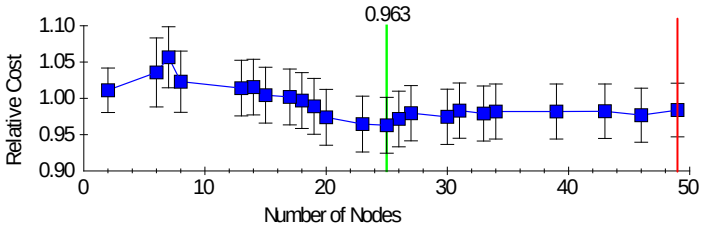
Hasil penyekatan rekursif secara biner dari data pengamatan yang digunakan akan menghasilkan pohon klasifikasi yang berukuran relatif besar dan tingkat kedalaman yang tinggi. Pohon klasifikasi tersebut disebut sebagai pohon klasifikasi maksimal yang ditunjukkan pada Gambar 4.14.



Gambar 4.14 Topologi Pohon Maksimal untuk Klasifikasi Curah Hujan pada Stasiun Pondok Betung Sebelum SMOTE

4.3.3.2 Pemangkasan Pohon Klasifikasi Maksimal (*Prunning*)

Guna mempermudah proses analisis, pohon klasifikasi maksimal yang dihasilkan kemudian dilakukan pemangkasan secara iteratif berdasarkan kriteria *cross-validated relative cost*. Setiap hasil pemangkasan memiliki nilai *relative cost* tertentu, sehingga kemudian dipilih hasil pemangkasan dengan nilai *relative cost* yang minimum.



Gambar 4.15 Plot *Relative Cost* Klasifikasi Curah Hujan Stasiun Pondok Betung Sebelum SMOTE

Berdasarkan Gambar 4.15, garis hijau menunjukkan pohon klasifikasi optimal sedangkan garis merah menujukkan klasifikasi maksimal. Pohon klasifikasi maksimal yang terbentuk terdiri dari 49 *terminal nodes* dan *relative cost* sebesar 0,984±0,037 yang dapat dilihat pada Tabel 4.29. Pemangkasan pohon dilakukan secara iteratif berdasarkan *cross-validated relative cost* yang minimum. Tabel 4.29 menunjukkan bahwa nilai *cross-validated relative cost* yang minimum adalah pada saat jumlah pohon 11

dan *terminal nodes* 25 sehingga dapat dikatakan bahwa pohon klasifikasi optimal yang terbentuk terdiri dari 25 *terminal nodes*.

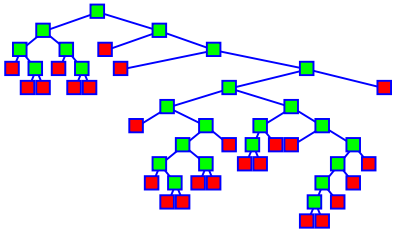
Tabel 4.29 Pembentukan Pohon Klasifikasi Stasiun Pondok Betung Sebelum SMOTE

Tree Number	*Terminal Nodes*	*Cross-validated Relative Cost*	*Resubsitution Relative Cost*
1	49	0,984 ± 0,037	0,053
11*	25	0,963 ± 0,038	0,137
15	18	0,997 ± 0,038	0,204
16	17	1,002 ± 0,038	0,221
17	15	1,004 ± 0,038	0,355
18	14	1,015 ± 0,038	0,273
19	13	1,014 ± 0,038	0,299
20	8	1,023 ± 0,042	0,447
21	7	1,057 ± 0,042	0,491
22	6	1,036 ± 0,048	0,539

*Pohon Klasifikasi Optimal

4.3.3.3 Pemilihan Pohon Klasifikasi Optimal

Hasil pemangkasan yang diperoleh dari Gambar 4.15 selanjutnya digunakan untuk memilih pohon klasifikasi yang optimal. Pohon klasifikasi optimal dengan jumlah *terminal nodes* 25, *cross-validated relative cost* sebesar 0,963 ± 0,038.



Gambar 4.16 Topologi Pohon Optimal untuk Klasifikasi Curah Hujan pada Stasiun Pondok Betung Sebelum SMOTE

Pembentukkan pohon klasifikasi optimal dipengaruhi oleh 35 variabel prediktor. Akan tetapi, urutan variabel terpenting dalam pohon klasifikasi optimal adalah PCustar, PCtscrn, PCtpan,

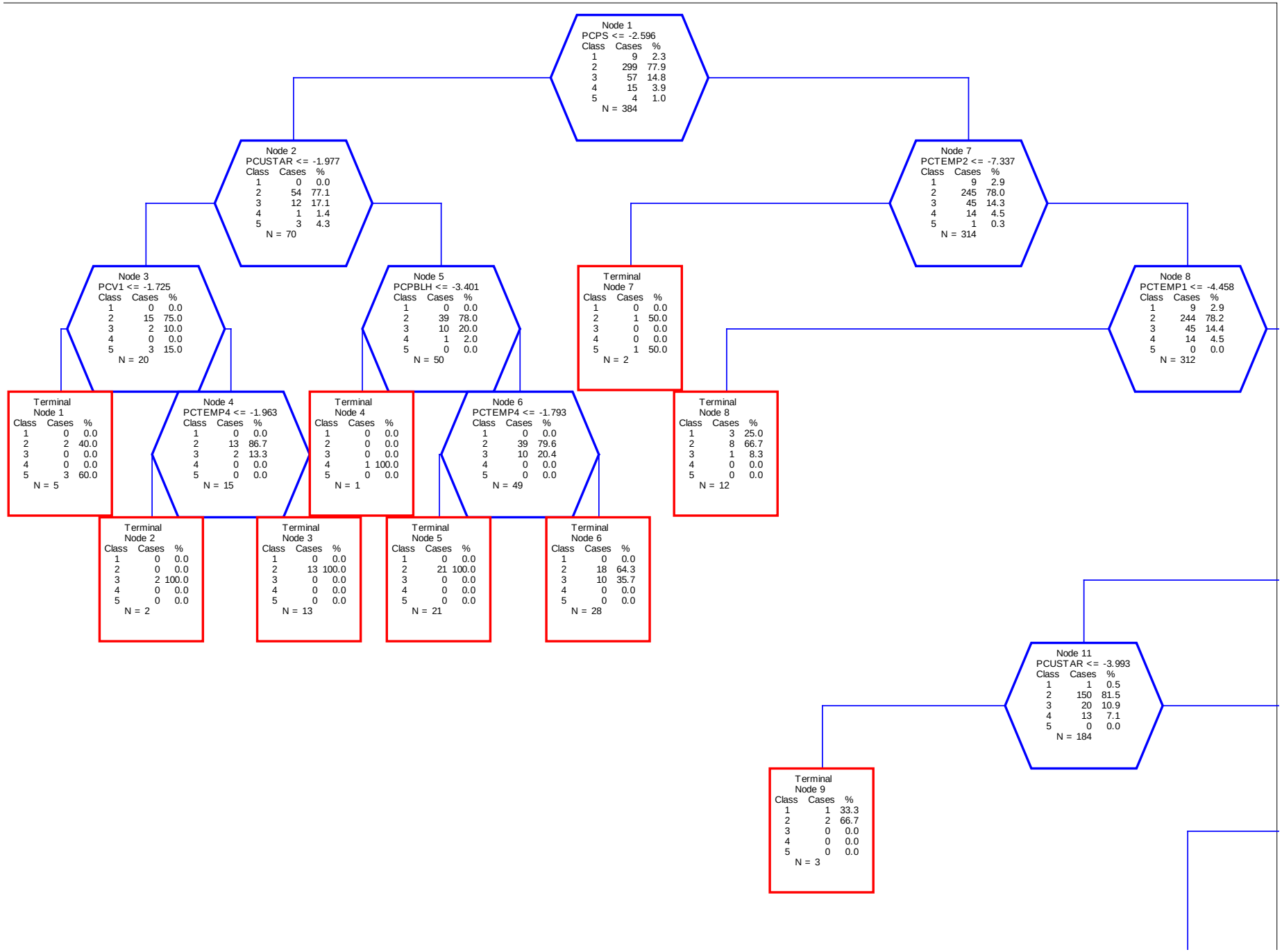
PCtemp1, PCtemp2, PCv1 dan seterusnya yang ditunjukkan pada Tabel 4.30.

Tabel 4.30 Variabel Penting Pembentukan Pohon Klasifikasi Optimal Stasiun Pondok Betung Sebelum SMOTE

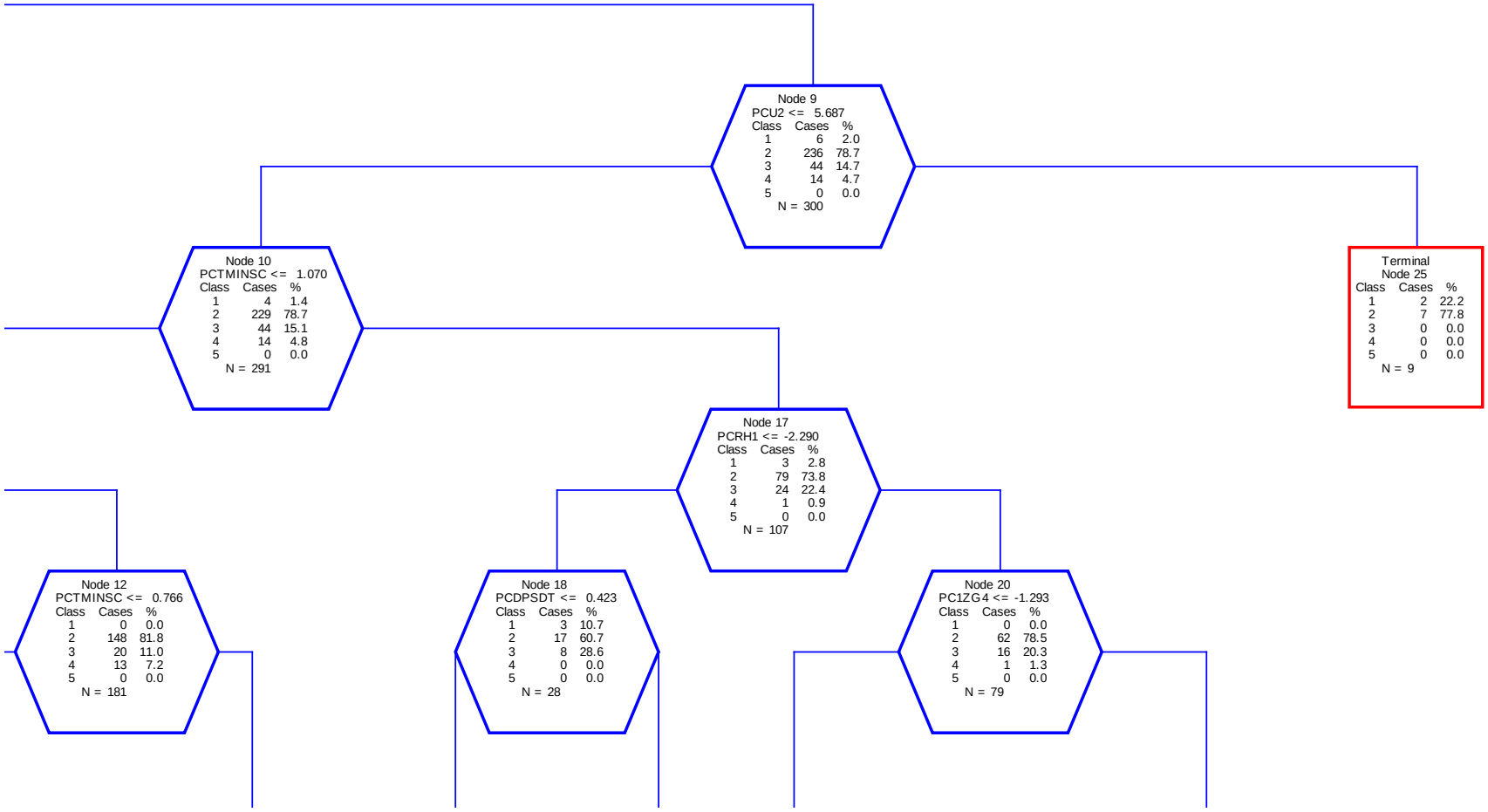
<table>
<tr><th>Variabel</th><th colspan="2">Skor Variabel</th></tr>
<tr><td>Pcustar</td><td>100</td><td>||||||||||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCtscrn</td><td>82,64</td><td>|||||||||||||||||||||||||||||||||||</td></tr>
<tr><td>Pctpan</td><td>77,98</td><td>|||||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCtemp1</td><td>72,85</td><td>||||||||||||||||||||||||||||||</td></tr>
<tr><td>PCtemp2</td><td>62,25</td><td>||||||||||||||||||||||||||</td></tr>
<tr><td>PCv1</td><td>60,31</td><td>|||||||||||||||||||||||||</td></tr>
</table>

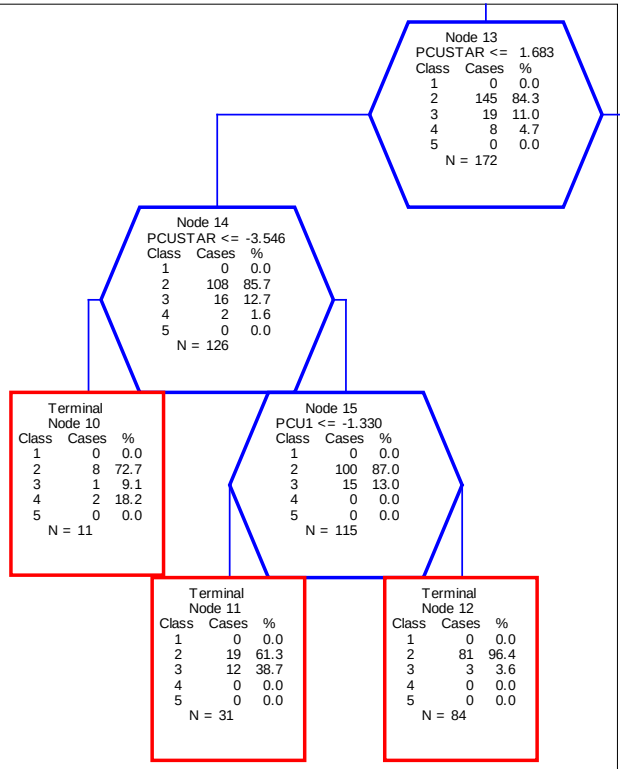
Walaupun variabel utama PCustar mempunyai skor 100, namun pada struktur pohon optimal ditunjukkan bahwa variabel yang menjadi pemilah *node* 1 adalah variabel PCps. Dimana jika nilai PCps ≤ -2,596 akan dipilah menjadi *node* kiri. Sedangkan jika nilai PCps > -2,596 akan dipilah menjadi *node* kanan. Gambar 4.17 merupakan visualisasi struktur pohon klasifikasi optimal.

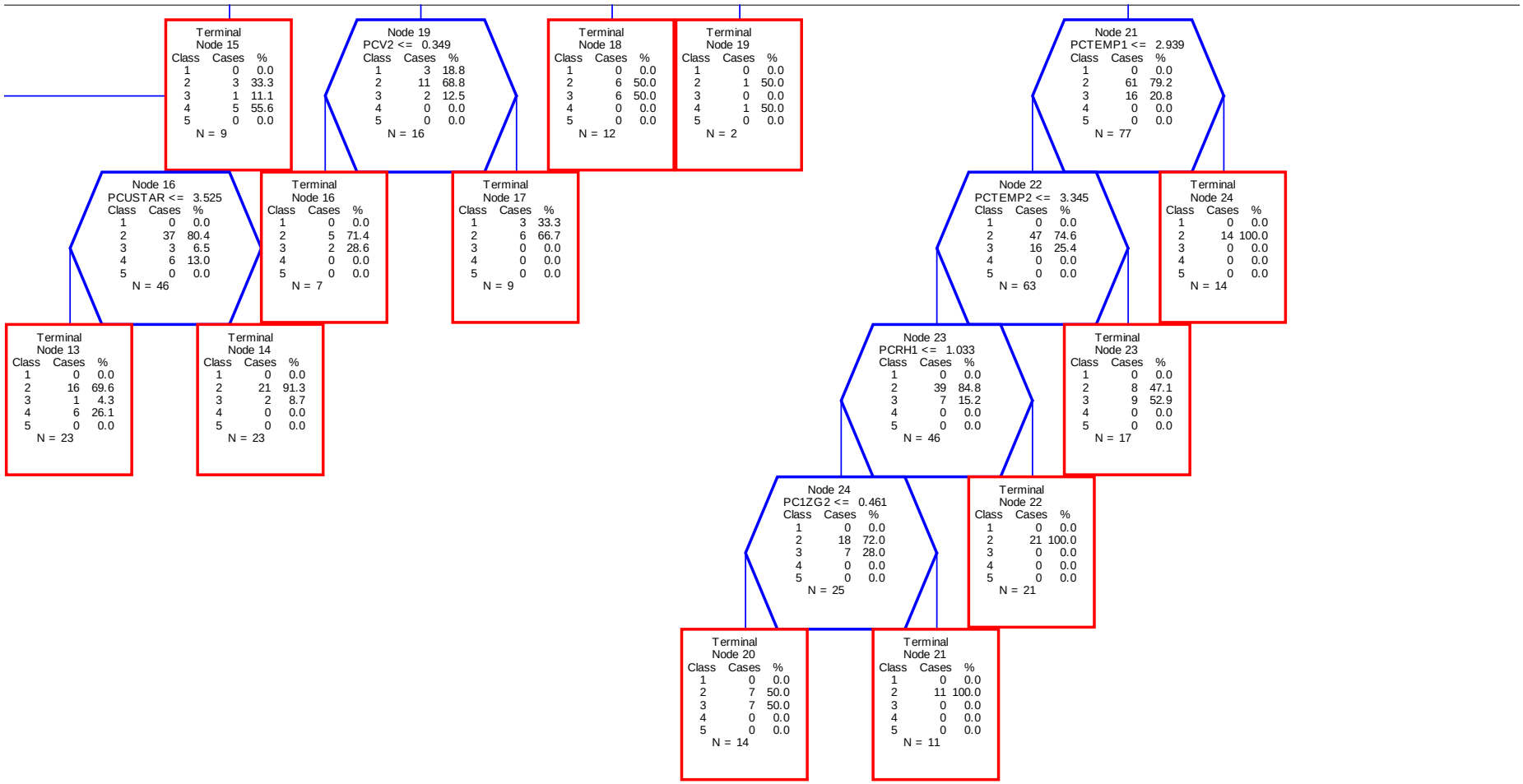
Suatu *node* akan terus dipilah menjadi *node* anak baru (kiri dan kanan) sesuai prosedur *binary recursive partitioning*, sampai *node* tersebut telah dianggap memiliki anggota yang homogen atau jika *node* tersebut hanya memiliki 1 anggota pengamatan maka *node* akan menjadi *node* terminal dan tidak akan dipilah lagi. Pohon klasifikasi optimal yang terbentuk terdiri atas 25 *node* terminal seperti pada Gambar 4.17. Masing-masing *node* terminal tersebut memiliki karakteristik tertentu dan diprediksi sebagai kelas variabel respon tertentu sesuai dengan label kelas yang diberikan. Berdasarkan hasil penelusuran 25 terminal *node* tersebut, Tabel 4.31 menampilkan rangkuman pengklasifikasian curah hujan menurut indikasi kesamaan label kelas setiap *node* terminal.



Gambar 4.17 Split Plot Pohon Optimal Stasiun Pondok Betung Sebelum SMOTE


Node 9
PCU2 <= 5.687
Class Cases %
1 6 2.0
2 236 78.7
3 44 14.7
4 14 4.7
5 0 0.0
N = 300
Node 10
PCTMINSC <= 1.070
Class Cases %
1 4 1.4
2 229 78.7
3 44 15.1
4 14 4.8
5 0 0.0
N = 291
Terminal
Node 25
Class Cases %
1 2 22.2
2 7 77.8
3 0 0.0
4 0 0.0
5 0 0.0
N = 9
Node 17
PCRH1 <= -2.290
Class Cases %
1 3 2.8
2 79 73.8
3 24 22.4
4 1 0.9
5 0 0.0
N = 107
Node 12
PCTMINSC <= 0.766
Class Cases %
1 0 0.0
2 148 81.8
3 20 11.0
4 13 7.2
5 0 0.0
N = 181
Node 18
PCDPSDT <= 0.423
Class Cases %
1 3 10.7
2 17 60.7
3 8 28.6
4 0 0.0
5 0 0.0
N = 28
Node 20
PC1ZG4 <= -1.293
Class Cases %
1 0 0.0
2 62 78.5
3 16 20.3
4 1 1.3
5 0 0.0
N = 79


Node 13
PCUSTAR <= 1.683
Class Cases %
1 0 0.0
2 145 84.3
3 19 11.0
4 8 4.7
5 0 0.0
N = 172
Node 14
PCUSTAR <= -3.546
Class Cases %
1 0 0.0
2 108 85.7
3 16 12.7
4 2 1.6
5 0 0.0
N = 126
Terminal
Node 10
Class Cases %
1 0 0.0
2 8 72.7
3 1 9.1
4 2 18.2
5 0 0.0
N = 11
Node 15
PCU1 <= -1.330
Class Cases %
1 0 0.0
2 100 87.0
3 15 13.0
4 0 0.0
5 0 0.0
N = 115
Terminal
Node 11
Class Cases %
1 0 0.0
2 19 61.3
3 12 38.7
4 0 0.0
5 0 0.0
N = 31
Terminal
Node 12
Class Cases %
1 0 0.0
2 81 96.4
3 3 3.6
4 0 0.0
5 0 0.0
N = 84


Terminal Node 15
Class Cases %
1 0 0.0
2 3 33.3
3 1 11.1
4 5 55.6
5 0 0.0
N = 9
Node 19
PCV2 <= 0.349
Class Cases %
1 3 18.8
2 11 68.8
3 2 12.5
4 0 0.0
5 0 0.0
N = 16
Terminal Node 18
Class Cases %
1 0 0.0
2 6 50.0
3 6 50.0
4 0 0.0
5 0 0.0
N = 12
Terminal Node 19
Class Cases %
1 0 0.0
2 1 50.0
3 0 0.0
4 1 50.0
5 0 0.0
N = 2
Node 21
PCTEMP1 <= 2.939
Class Cases %
1 0 0.0
2 61 79.2
3 16 20.8
4 0 0.0
5 0 0.0
N = 77
Node 16
PCUSTAR <= 3.525
Class Cases %
1 0 0.0
2 37 80.4
3 3 6.5
4 6 13.0
5 0 0.0
N = 46
Terminal Node 16
Class Cases %
1 0 0.0
2 5 71.4
3 2 28.6
4 0 0.0
5 0 0.0
N = 7
Terminal Node 17
Class Cases %
1 3 33.3
2 6 66.7
3 0 0.0
4 0 0.0
5 0 0.0
N = 9
Node 22
PCTEMP2 <= 3.345
Class Cases %
1 0 0.0
2 47 74.6
3 16 25.4
4 0 0.0
5 0 0.0
N = 63
Terminal Node 24
Class Cases %
1 0 0.0
2 14 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 14
Terminal Node 13
Class Cases %
1 0 0.0
2 16 69.6
3 1 4.3
4 6 26.1
5 0 0.0
N = 23
Terminal Node 14
Class Cases %
1 0 0.0
2 21 91.3
3 2 8.7
4 0 0.0
5 0 0.0
N = 23
Node 23
PCRH1 <= 1.033
Class Cases %
1 0 0.0
2 39 84.8
3 7 15.2
4 0 0.0
5 0 0.0
N = 46
Terminal Node 23
Class Cases %
1 0 0.0
2 8 47.1
3 9 52.9
4 0 0.0
5 0 0.0
N = 17
Node 24
PC1ZG2 <= 0.461
Class Cases %
1 0 0.0
2 18 72.0
3 7 28.0
4 0 0.0
5 0 0.0
N = 25
Terminal Node 22
Class Cases %
1 0 0.0
2 21 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 21
Terminal Node 20
Class Cases %
1 0 0.0
2 7 50.0
3 7 50.0
4 0 0.0
5 0 0.0
N = 14
Terminal Node 21
Class Cases %
1 0 0.0
2 11 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 11

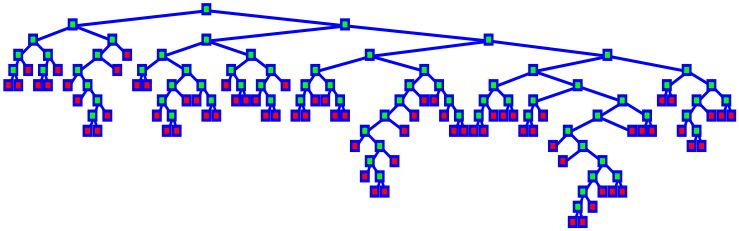
Tabel 4.31 Kelas Curah Hujan Stasiun Pondok Betung pada Masing-Masing Terminal *Node* Sebelum SMOTE

Kelas	*Terminal Node*	Persentase	*Terminal Node*	Persentase
1	8	88,3	17	94,3
	9	94,3	25	90,5
2	3	100	21	100
	5	100	22	100
	12	83,7	24	100
	14	66,7	21	100
3	2	100	18	84
	6	74,5	20	84
	11	76,8	23	85,5
	16	67,7		
4	4	100	15	92,4
	10	75,1	19	95,2
	13	84,9		
5	1	99,1	7	98,7

Dari Tabel 4.31 dapat lihat bahwa kelas 2 merupakan kelas klasifikasi yang paling banyak terbentuk. Hal ini disebabkan distribusi data curah hujan tidak *balance* antara cerah berawan, hujan ringan, hujan sedang, hujan lebat, dan hujan lebat sekali. Pada pengamatan stasiun Pondok Betung data didominasi dengan hujan ringan (kelas 2) dan jumlah data paling rendah berada pada kejadian hujan lebat sekali (kelas 5). Sehingga pada Tabel 4.31, kelas 5 hanya terdapat pada 2 terminal *node*. Maka dari itu, pada stasiun Pondok Betung akan dicobakan proses SMOTE untuk mengatasi masalah *imbalance* data.

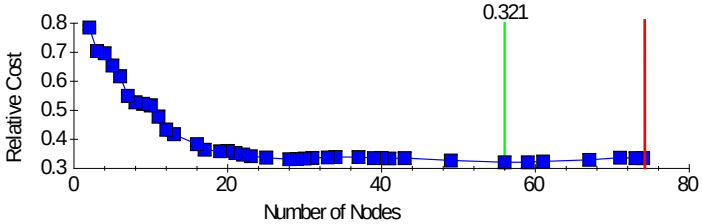
Proses SMOTE pada stasiun Pondok Betung dilakukan sebanyak 12 iterasi dan menambah jumlah data pengamatan menjadi 733 pengamatan. Kemudian data pengamatan tersebut digunakan untuk membangun model klasifikasi pohon yang baru.

Pohon maksimal yang dihasilkan memiliki 74 terminal *node* dengan kedalaman sebesar 14 tingkatan. Variabel yang menjadi pemilah utama adalah variabel PCmixr2 dengan skor 100. Topologi pohon maksimal ditampilkan pada Gambar 4.18.



Gambar 4.18 Topologi Pohon Maksimal Stasiun Pondok Betung Setelah SMOTE

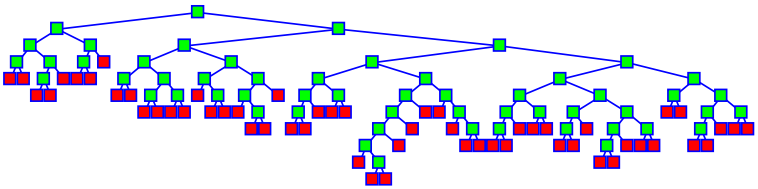
Gambar 4.19 menampilkan adanya perbedaan nilai *relative cost* yang dihasilkan oleh pohon klasifikasi maksimal dengan pohon klasifikasi yang dianggap optimal. Pohon klasifikasi maksimal ditunjukkan oleh garis berwarna merah sedangkan pohon klasifikasi optimal ditunjukkan oleh garis berwarna hijau.



Gambar 4.19 Plot *Relative Cost* Klasifikasi Curah Hujan Stasiun Pondok Betung Setelah SMOTE

Berdasarkan Gambar 4.19, pohon klasifikasi maksimal yang terbentuk terdiri dari 74 terminal *nodes* dan *relative cost* sebesar 0,335 ± 0,021. Pemangkasan pohon dilakukan secara iteratif berdasarkan *cross validated relative cost* yang minimum. Nilai *cross validated relative cost* yang minimum adalah 0,321 ± 0,020 pada saat *terminal nodes* sebanyak 56. Sehingga dapat dikatakan bahwa pohon klasifikasi optimal yang terbentuk terdiri

dari 56 *terminal nodes*. Karena nilai *relative cost* pohon klasifikasi optimal lebih kecil maka pohon klasifikasi optimal dipilih sebagai pohon yang layak untuk pohon klasifikasi curah hujan pada stasiun pengamatan Pondok Betung.



Gambar 4.20 Topologi Pohon Optimal Stasiun Pondok Betung Setelah SMOTE

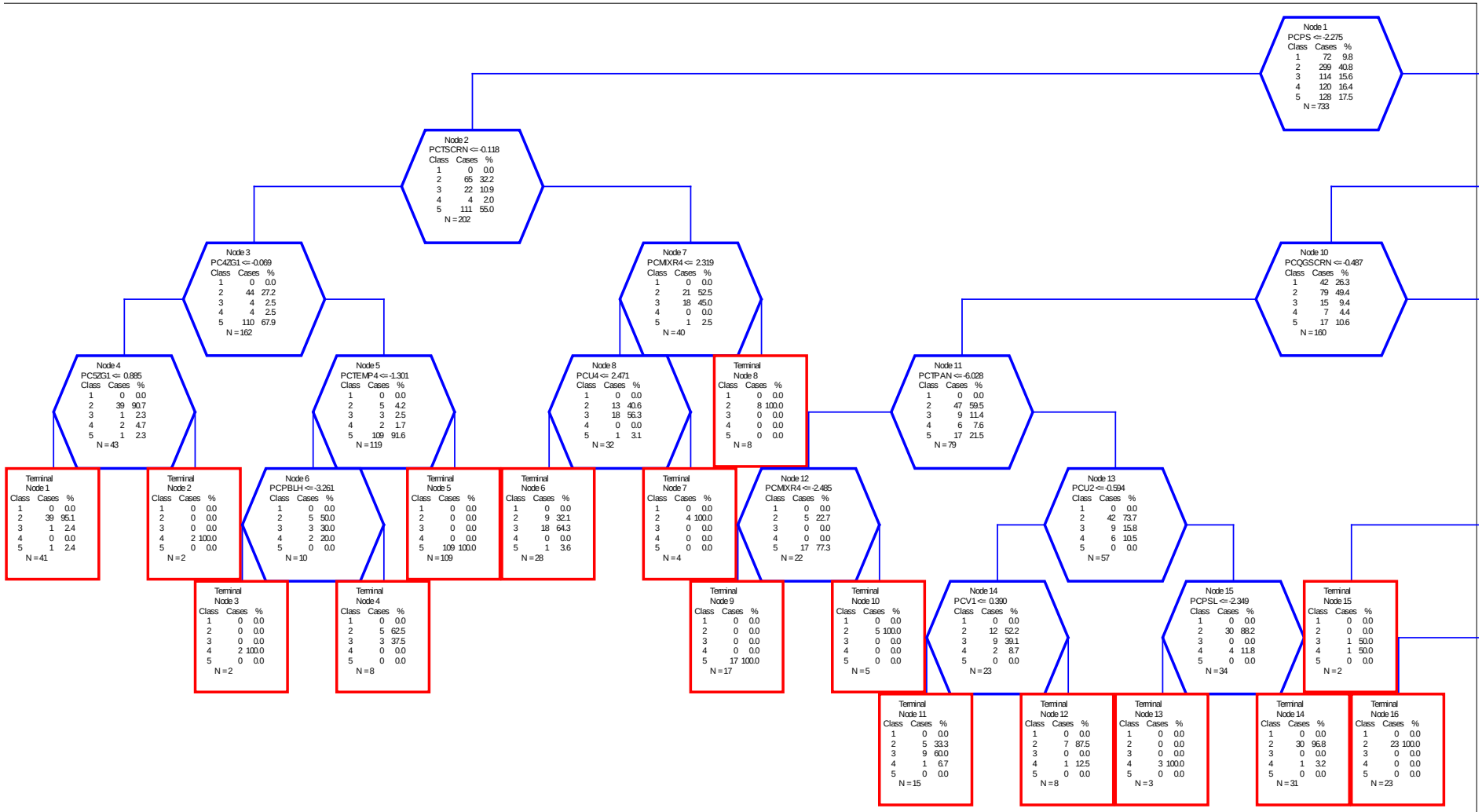
Berdasarkan topologi pohon klasifikasi optimal, diketahui bahwa PCmixr2 merupakan variabel pemilah yang utama dan paling penting dalam menentukan klasifikasi curah hujan di stasiun pengamatan Kemayoran. Pada Tabel 4.32, skor variabel PCmixr2 adalah 100 karena mampu memberikan nilai penurunan keheterogenan tertinggi pada *node* utama. Selain itu ada 32 variabel lain yang juga berkontribusi dalam pembentukan pohon klasifikasi optimal, hasil selengkapnya disajikan dalam Lampiran 21.

Tabel 4.32 Variabel Penting Pembentukan Pohon Klasifikasi Optimal Stasiun Pondok Betung Setelah SMOTE

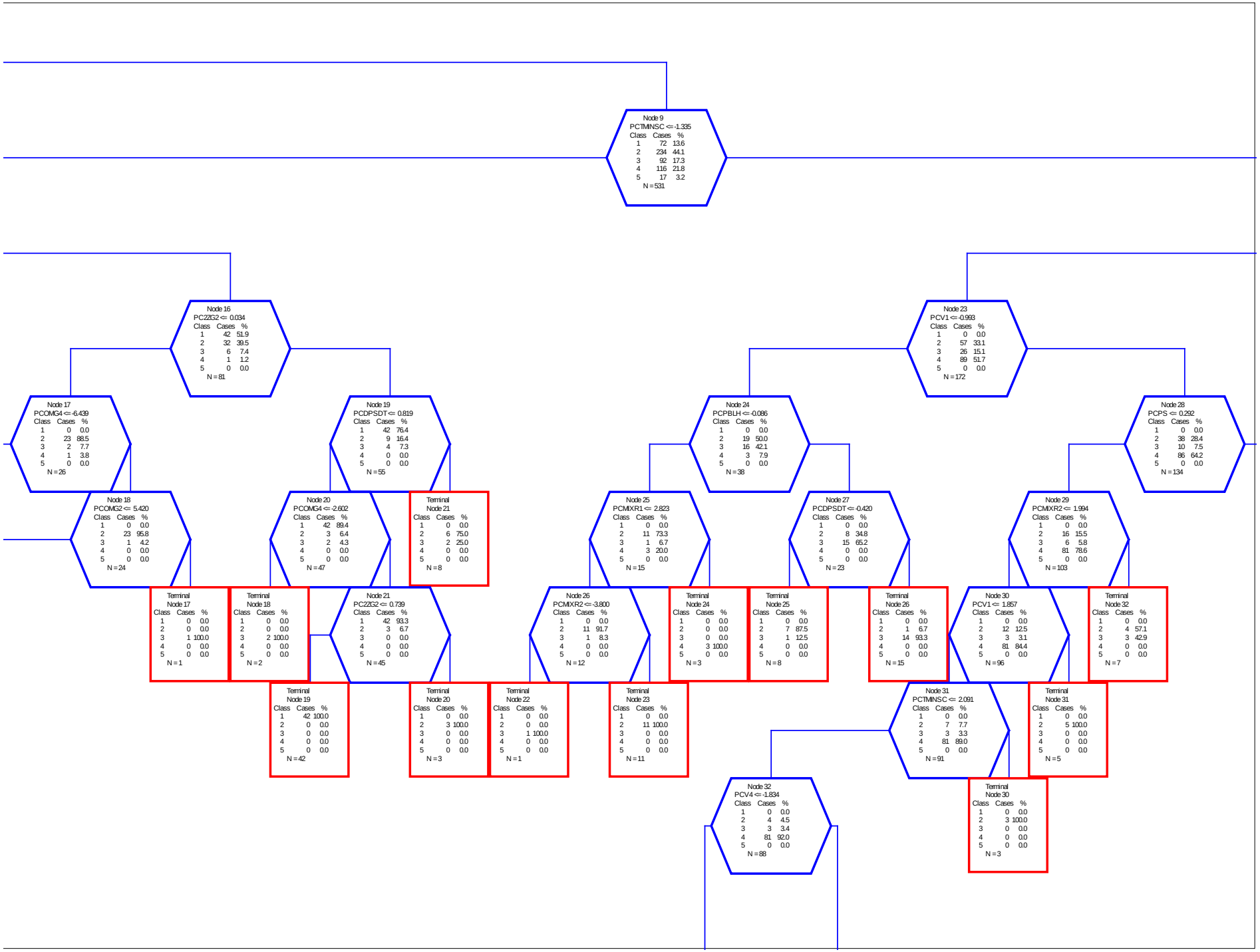
Variabel	Skor Variabel	
PCmixr2	100	||||||||||||||||||||||||||||||||||||||||||
PCps	85,22	||||||||||||||||||||||||||||||||||||
Pctpan	74,66	|||||||||||||||||||||||||||||||
PCpsl	69,67	|||||||||||||||||||||||||||||
PCrh4	67,87	||||||||||||||||||||||||||||

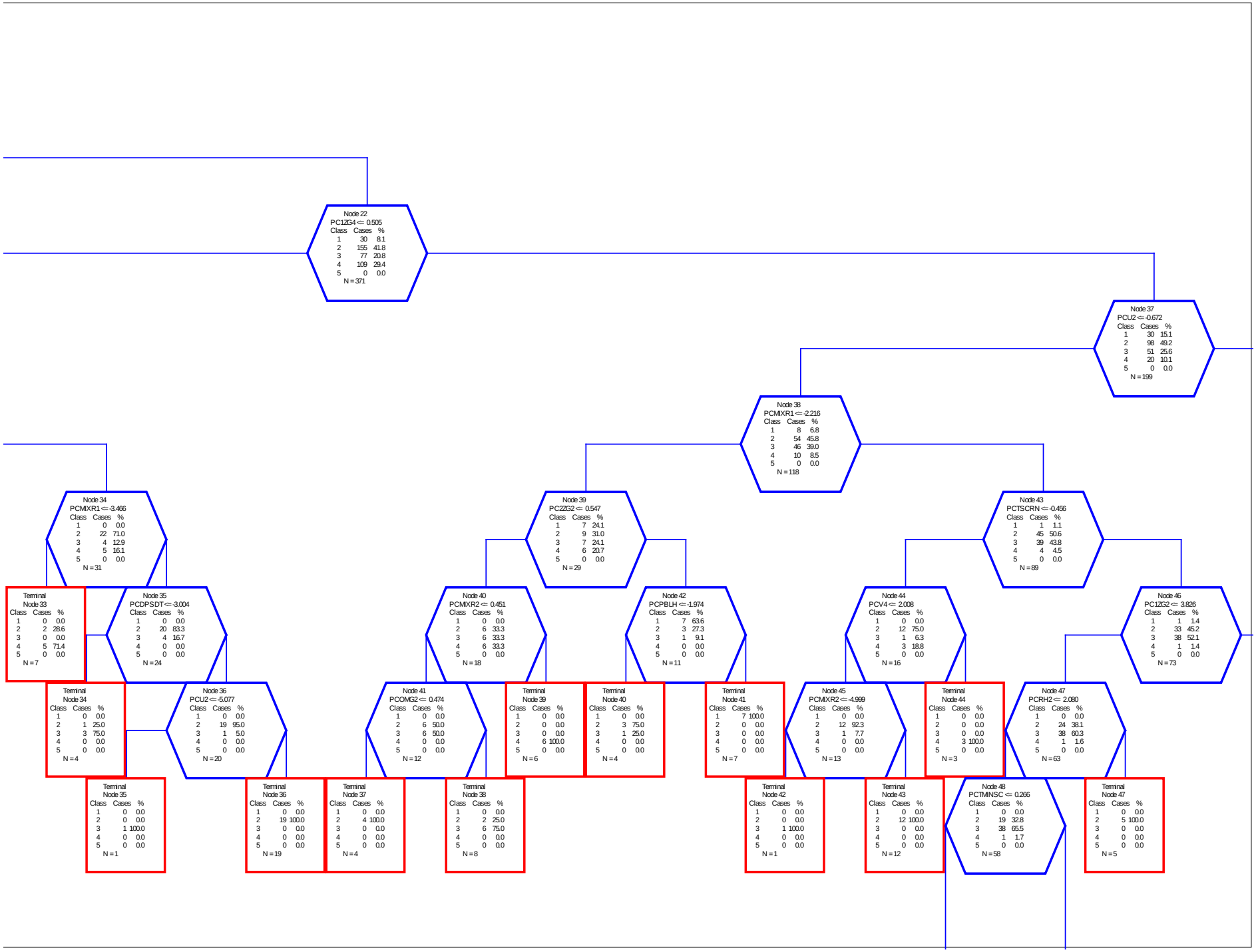
Pada struktur pohon optimal, dapat dilihat bahwa yang menjadi pemilah *node* 1 adalah variabel PCps. Variabel utama PCps memilah *node* utama menjadi *node* kanan dan kiri dengan ketentuan nilai PCps ≤ -2,275 akan dipilah menjadi *node* kiri. Sedangkan jika nilai PCps > -2,275 akan dipilah menjadi *node* kanan. Gambar 4.21 merupakan visualisasi struktur pohon klasifikasi optimal.

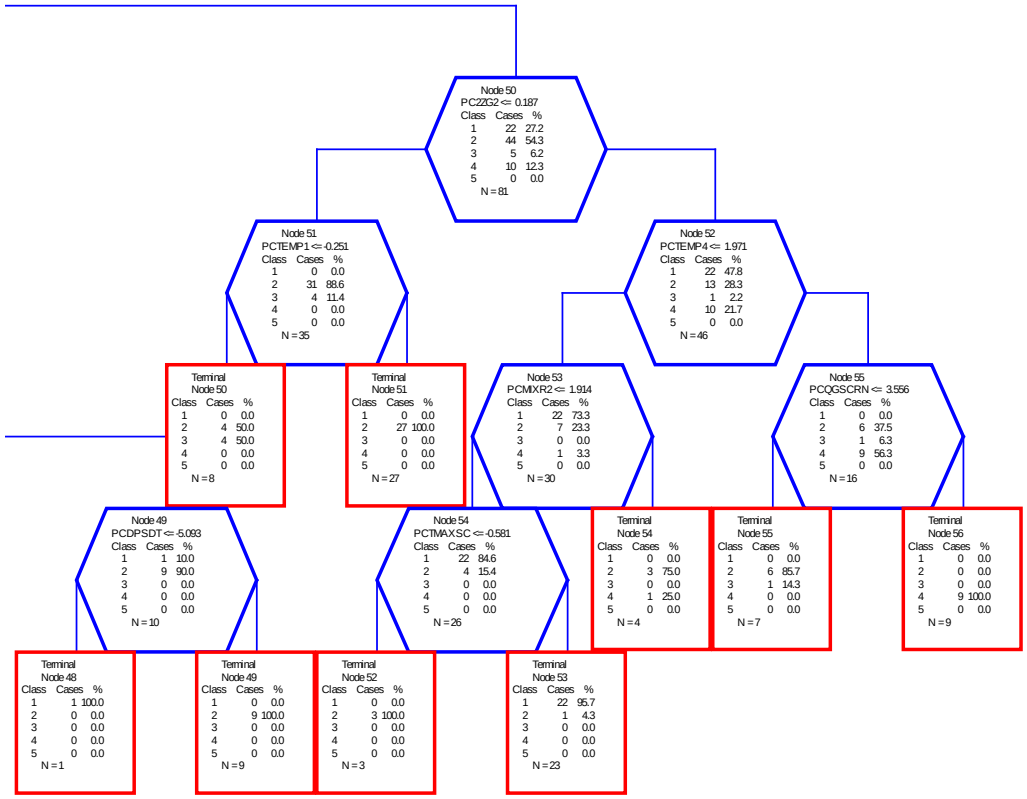
(Halaman ini sengaja dikosongkan)

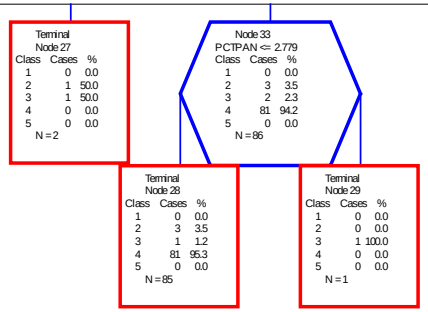


Gambar 4.21 Split Plot Pohon Optimal Stasiun Pondok Betung Setelah SMOTE


Node 9
PCTMINSC <= -1.335
Class Cases %
1 72 13.6
2 234 44.1
3 92 17.3
4 116 21.8
5 17 3.2
N = 531
Node 16
PC22G2 <= 0.034
Class Cases %
1 42 51.9
2 32 39.5
3 6 7.4
4 1 1.2
5 0 0.0
N = 81
Node 23
PCV1 <= -0.993
Class Cases %
1 0 0.0
2 57 33.1
3 26 15.1
4 89 51.7
5 0 0.0
N = 172
Node 17
PCOMG4 <= -6.439
Class Cases %
1 0 0.0
2 23 88.5
3 2 7.7
4 1 3.8
5 0 0.0
N = 26
Node 19
PCDPSDT <= 0.819
Class Cases %
1 42 76.4
2 9 16.4
3 4 7.3
4 0 0.0
5 0 0.0
N = 55
Node 24
PCPBLH <= -0.086
Class Cases %
1 0 0.0
2 19 50.0
3 16 42.1
4 3 7.9
5 0 0.0
N = 38
Node 28
PCPS <= 0.292
Class Cases %
1 0 0.0
2 38 28.4
3 10 7.5
4 86 64.2
5 0 0.0
N = 134
Node 18
PCOMG2 <= 5.420
Class Cases %
1 0 0.0
2 23 95.8
3 1 4.2
4 0 0.0
5 0 0.0
N = 24
Node 20
PCOMG4 <= -2.602
Class Cases %
1 42 89.4
2 3 6.4
3 2 4.3
4 0 0.0
5 0 0.0
N = 47
Terminal
Node 21
Class Cases %
1 0 0.0
2 6 75.0
3 2 25.0
4 0 0.0
5 0 0.0
N = 8
Node 25
PCMXR1 <= 2.823
Class Cases %
1 0 0.0
2 11 73.3
3 1 6.7
4 3 20.0
5 0 0.0
N = 15
Node 27
PCDPSDT <= -0.420
Class Cases %
1 0 0.0
2 8 34.8
3 15 65.2
4 0 0.0
5 0 0.0
N = 23
Node 29
PCMXR2 <= 1.994
Class Cases %
1 0 0.0
2 16 15.5
3 6 5.8
4 81 78.6
5 0 0.0
N = 103
Terminal
Node 17
Class Cases %
1 0 0.0
2 0 0.0
3 1 100.0
4 0 0.0
5 0 0.0
N = 1
Terminal
Node 18
Class Cases %
1 0 0.0
2 0 0.0
3 2 100.0
4 0 0.0
5 0 0.0
N = 2
Node 21
PC22G2 <= 0.739
Class Cases %
1 42 93.3
2 3 6.7
3 0 0.0
4 0 0.0
5 0 0.0
N = 45
Node 26
PCMXR2 <= -3.800
Class Cases %
1 0 0.0
2 11 91.7
3 1 8.3
4 0 0.0
5 0 0.0
N = 12
Terminal
Node 24
Class Cases %
1 0 0.0
2 0 0.0
3 0 0.0
4 3 100.0
5 0 0.0
N = 3
Terminal
Node 25
Class Cases %
1 0 0.0
2 7 87.5
3 1 12.5
4 0 0.0
5 0 0.0
N = 8
Terminal
Node 26
Class Cases %
1 0 0.0
2 1 6.7
3 14 93.3
4 0 0.0
5 0 0.0
N = 15
Node 30
PCV1 <= 1.857
Class Cases %
1 0 0.0
2 12 12.5
3 3 3.1
4 81 84.4
5 0 0.0
N = 96
Terminal
Node 32
Class Cases %
1 0 0.0
2 4 57.1
3 3 42.9
4 0 0.0
5 0 0.0
N = 7
Terminal
Node 19
Class Cases %
1 42 100.0
2 0 0.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 42
Terminal
Node 20
Class Cases %
1 0 0.0
2 3 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 3
Terminal
Node 22
Class Cases %
1 0 0.0
2 0 0.0
3 1 100.0
4 0 0.0
5 0 0.0
N = 1
Terminal
Node 23
Class Cases %
1 0 0.0
2 11 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 11
Node 31
PCTMINSC <= 2.091
Class Cases %
1 0 0.0
2 7 7.7
3 3 3.3
4 81 89.0
5 0 0.0
N = 91
Terminal
Node 31
Class Cases %
1 0 0.0
2 5 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 5
Node 32
PCV4 <= -1.834
Class Cases %
1 0 0.0
2 4 4.5
3 3 3.4
4 81 92.0
5 0 0.0
N = 88
Terminal
Node 30
Class Cases %
1 0 0.0
2 3 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 3


Node 22
PC12G4 <= 0.505
Class Cases %
1 30 8.1
2 155 41.8
3 77 20.8
4 109 29.4
5 0 0.0
N = 371
Node 37
PCU2 <= -0.672
Class Cases %
1 30 15.1
2 98 49.2
3 51 25.6
4 20 10.1
5 0 0.0
N = 199
Node 38
PCMIXR1 <= -2.216
Class Cases %
1 8 6.8
2 54 45.8
3 46 39.0
4 10 8.5
5 0 0.0
N = 118
Node 34
PCMIXR1 <= -3.466
Class Cases %
1 0 0.0
2 22 71.0
3 4 12.9
4 5 16.1
5 0 0.0
N = 31
Node 39
PC22G2 <= 0.547
Class Cases %
1 7 24.1
2 9 31.0
3 7 24.1
4 6 20.7
5 0 0.0
N = 29
Node 43
PCTSCRN <= -0.456
Class Cases %
1 1 1.1
2 45 50.6
3 39 43.8
4 4 4.5
5 0 0.0
N = 89
Terminal
Node 33
Class Cases %
1 0 0.0
2 2 28.6
3 0 0.0
4 5 71.4
5 0 0.0
N = 7
Node 35
PCDPSDT <= -3.004
Class Cases %
1 0 0.0
2 20 83.3
3 4 16.7
4 0 0.0
5 0 0.0
N = 24
Node 40
PCMIXR2 <= 0.451
Class Cases %
1 0 0.0
2 6 33.3
3 6 33.3
4 6 33.3
5 0 0.0
N = 18
Node 42
PCPBLH <= -1.974
Class Cases %
1 7 63.6
2 3 27.3
3 1 9.1
4 0 0.0
5 0 0.0
N = 11
Node 44
PCV4 <= 2.008
Class Cases %
1 0 0.0
2 12 75.0
3 1 6.3
4 3 18.8
5 0 0.0
N = 16
Node 46
PC12G2 <= 3.826
Class Cases %
1 1 1.4
2 33 45.2
3 38 52.1
4 1 1.4
5 0 0.0
N = 73
Terminal
Node 34
Class Cases %
1 0 0.0
2 1 25.0
3 3 75.0
4 0 0.0
5 0 0.0
N = 4
Node 36
PCU2 <= -5.077
Class Cases %
1 0 0.0
2 19 95.0
3 1 5.0
4 0 0.0
5 0 0.0
N = 20
Node 41
PCOMG2 <= 0.474
Class Cases %
1 0 0.0
2 6 50.0
3 6 50.0
4 0 0.0
5 0 0.0
N = 12
Terminal
Node 39
Class Cases %
1 0 0.0
2 0 0.0
3 0 0.0
4 6 100.0
5 0 0.0
N = 6
Terminal
Node 40
Class Cases %
1 0 0.0
2 3 75.0
3 1 25.0
4 0 0.0
5 0 0.0
N = 4
Terminal
Node 41
Class Cases %
1 7 100.0
2 0 0.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 7
Node 45
PCMIXR2 <= -4.999
Class Cases %
1 0 0.0
2 12 92.3
3 1 7.7
4 0 0.0
5 0 0.0
N = 13
Terminal
Node 44
Class Cases %
1 0 0.0
2 0 0.0
3 0 0.0
4 3 100.0
5 0 0.0
N = 3
Node 47
PCRH2 <= 2.080
Class Cases %
1 0 0.0
2 24 38.1
3 38 60.3
4 1 1.6
5 0 0.0
N = 63
Terminal
Node 35
Class Cases %
1 0 0.0
2 0 0.0
3 1 100.0
4 0 0.0
5 0 0.0
N = 1
Terminal
Node 36
Class Cases %
1 0 0.0
2 19 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 19
Terminal
Node 37
Class Cases %
1 0 0.0
2 4 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 4
Terminal
Node 38
Class Cases %
1 0 0.0
2 2 25.0
3 6 75.0
4 0 0.0
5 0 0.0
N = 8
Terminal
Node 42
Class Cases %
1 0 0.0
2 0 0.0
3 1 100.0
4 0 0.0
5 0 0.0
N = 1
Terminal
Node 43
Class Cases %
1 0 0.0
2 12 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 12
Node 48
PCTMINSC <= 0.266
Class Cases %
1 0 0.0
2 19 32.8
3 38 65.5
4 1 1.7
5 0 0.0
N = 58
Terminal
Node 47
Class Cases %
1 0 0.0
2 5 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 5


Node 50
PCZZGS2 <= 0.187
Class Cases %
1 22 27.2
2 44 54.3
3 5 6.2
4 10 12.3
5 0 0.0
N = 81
Node 51
PCTEMP1 <= -0.251
Class Cases %
1 0 0.0
2 31 88.6
3 4 11.4
4 0 0.0
5 0 0.0
N = 35
Node 52
PCTEMP4 <= 1.971
Class Cases %
1 22 47.8
2 13 28.3
3 1 2.2
4 10 21.7
5 0 0.0
N = 46
Terminal
Node 50
Class Cases %
1 0 0.0
2 4 50.0
3 4 50.0
4 0 0.0
5 0 0.0
N = 8
Terminal
Node 51
Class Cases %
1 0 0.0
2 27 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 27
Node 53
PCMIXR2 <= 1.914
Class Cases %
1 22 73.3
2 7 23.3
3 0 0.0
4 1 3.3
5 0 0.0
N = 30
Node 55
PCQGSCRN <= 3.556
Class Cases %
1 0 0.0
2 6 37.5
3 1 6.3
4 9 56.3
5 0 0.0
N = 16
Node 49
PCDPSDT <= -5.093
Class Cases %
1 1 10.0
2 9 90.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 10
Node 54
PCTMAXSC <= -0.581
Class Cases %
1 22 84.6
2 4 15.4
3 0 0.0
4 0 0.0
5 0 0.0
N = 26
Terminal
Node 54
Class Cases %
1 0 0.0
2 3 75.0
3 0 0.0
4 1 25.0
5 0 0.0
N = 4
Terminal
Node 55
Class Cases %
1 0 0.0
2 6 85.7
3 1 14.3
4 0 0.0
5 0 0.0
N = 7
Terminal
Node 56
Class Cases %
1 0 0.0
2 0 0.0
3 0 0.0
4 9 100.0
5 0 0.0
N = 9
Terminal
Node 48
Class Cases %
1 1 100.0
2 0 0.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 1
Terminal
Node 49
Class Cases %
1 0 0.0
2 9 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 9
Terminal
Node 52
Class Cases %
1 0 0.0
2 3 100.0
3 0 0.0
4 0 0.0
5 0 0.0
N = 3
Terminal
Node 53
Class Cases %
1 22 95.7
2 1 4.3
3 0 0.0
4 0 0.0
5 0 0.0
N = 23


Terminal
Node 27
Class Cases %
1 0 0.0
2 1 50.0
3 1 50.0
4 0 0.0
5 0 0.0
N = 2
Node 33
PCTPAN <= 2.779
Class Cases %
1 0 0.0
2 3 3.5
3 2 2.3
4 81 94.2
5 0 0.0
N = 86
Terminal
Node 28
Class Cases %
1 0 0.0
2 3 3.5
3 1 1.2
4 81 95.3
5 0 0.0
N = 85
Terminal
Node 29
Class Cases %
1 0 0.0
2 0 0.0
3 1 100.0
4 0 0.0
5 0 0.0
N = 1

Terminal
Node 45

Class	Cases	%
1	0	0.0
2	3	100.0
3	0	0.0
4	0	0.0
5	0	0.0

N = 3

Terminal
Node 46

Class	Cases	%
1	0	0.0
2	16	29.1
3	38	69.1
4	1	1.8
5	0	0.0

N = 55

Suatu *node* akan terus dipilah menjadi *node* anak baru (kiri dan kanan) sesuai prosedur *binary recursive partitioning*, sampai *node* tersebut telah dianggap memiliki anggota yang homogen atau jika *node* tersebut hanya memiliki 1 anggota pengamatan maka *node* akan menjadi *node* terminal dan tidak akan dipilah lagi. Pohon klasifikasi optimal yang terbentuk terdiri atas 56 *node* terminal seperti pada Gambar 4.21. Masing-masing *node* terminal tersebut memiliki karakteristik tertentu dan diprediksi sebagai kelas variabel respon tertentu sesuai dengan label kelas yang diberikan. Berdasarkan hasil penelusuran 56 terminal *node* tersebut, Tabel 4.33 menampilkan rangkuman pengklasifikasian curah hujan menurut indikasi kesamaan label kelas setiap *node* terminal.

Tabel 4.33 Kelas Curah Hujan Stasiun Pondok Betung pada Masing-Masing Terminal *Node* Setelah SMOTE

Kelas	***Terminal Node***	**Persentase**	***Terminal Node***	**Persentase**
1	19	100	48	100
	41	100	53	98,9
2	1	88,7	31	100
	7	100	36	100
	8	100	37	100
	10	100	40	53,4
	12	73,7	43	100
	14	92,3	45	100
	16	100	47	100
	20	100	49	100
	21	53,4	51	100
	23	100	52	100
	25	72,7	54	54,6
	30	100	55	69,6
3	4	61,1	29	100

Tabel 4.33 (Lanjutan) Kelas Curah Hujan Stasiun Pondok Betung pada Masing-Masing Terminal *Node* Setelah SMOTE

Kelas	*Terminal Node*	Persentase	*Terminal Node*	Persentase
3	6	80,6	32	66,3
	11	75,9	34	88,7
	15	51,3	35	100
	17	100	38	88,7
	18	100	42	100
	22	100	46	84,4
	26	97,3	50	72,4
	27	72,4		
4	2	100	33	86,2
	3	100	39	100
	13	100	44	100
	24	100	56	100
	28	97,3		
5	5	100		
	9	100		

Dari Tabel 4.33 dapat diketahui walaupun sudah dilakukan proses SMOTE untuk mengatasi data *imbalance*, tetapi dari 56 terminal *node* yang terbentuk masih cenderung pada kelas 2. Karena pada setelah dilakukan proses SMOTE, data pada kelas 2 masih tetap paling tinggi dibandingkan kelas lainnya. Secara keseluruhan dapat diketahui bahwa terdapat 72 pengamatan dalam kelas cerah berawan, 245 pengamatan masuk dalam kelas hujan ringan, 107 pengamatan termasuk dalam kelas hujan sedang, 114 pengamatan masuk dalam kelas hujan lebat dan 126 pengamatan yang termasuk dalam kelas hujan lebat sekali.

Penelusuran struktur pohon klasifikasi optimal terhadap *node* terminal dapat memberikan informasi tentang karakteristik kelas *node* terminal dengan persentase tertinggi untuk masing-

masing kelas. Karakteristik kelas curah hujan pada masing-masing *node* terminal disajikan pada Tabel 4.34.

Tabel 4.34 Karakteristik Kelas Curah Hujan Stasiun Pondok Betung Setelah SMOTE

Kelas	Karakteristik
Cerah Berawan (1)	PCps > -2,274 PCtminscr ≤ -1,133 PCqgscrn > -0,486 PCdpsdt ≤ 0,819 PComega4 > -2,601 PC2zg2 > 0,033 PC2zg2 ≤ 0,738
Hujan Ringan (2)	PCps ≤ -2,274 PCtscrn > -0,118 PCmixr4 ≤ 2,318 PCu4 > 2,478
Hujan Sedang (3)	PCps > -2,274 PCminscr ≤ -1,334 PCqgscrn > -0,486 PC2zg2 ≤ 0,033 PComega4 > -6,439 PComega2 > 5,420
Hujan Lebat (4)	PCps ≤ -2,274 PCtscrn ≤ -0,118 PC4zg1 ≤ -0,068 PC5zg1 > 0,885
Hujan Lebat Sekali (5)	PCps ≤ -2,274 PCtscrn ≤ -0,118 PC4zg1 > -0,068 PCtemp4 > -1,3

4.3.3.4 Hasil Ketepatan Klasifikasi Klasifikasi Pohon

Tingkat keakuratan hasil klasifikasi pohon optimal yang dihasilkan dari data *learning* dapat dihitung berdasarkan Tabel 4.35.

Tabel 4.35 Klasifikasi Curah Hujan Data *Learning* pada Pohon Optimal Stasiun Pondok Betung Sebelum SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	1	3	4	1	0	11	8
2	23	155	75	43	3	52	144
3	5	24	22	4	2	39	35
4	0	9	3	2	1	0	13
5	0	2	2	0	0	0	4

Berdasarkan Tabel 4.35, kesalahan klasifikasi kelas pengamatan terjadi pada seluruh kelas, dimana kelas 4 dan 5 menghasilkan ketepatan klasifikasi sebesar 0%. Artinya tidak ada 1 pun data pengamatan kelas 4 dan kelas 5 yang diklasifikasikan secara tepat. Ketepatan klasifikasi terbesar terjadi pada kelas 2 (hujan ringan) dengan persentase sebesar 52%.

Menggunakan informasi pada Tabel 4.35, maka ketepatan klasifikasi data *learning* dapat dihitung sebagai berikut:

$$1 - APER = \left(1 - \frac{8 + 144 + 35 + 13 + 4}{384}\right) \times 100\% = 46{,}88\%$$

Hasil perhitungan ketepatan klasifikasi data *learning* sebesar 46,88 persen. Artinya pohon klasifikasi optimal mampu mengklasifikasikan pengamatan curah hujan kedalam kelas kategori hujan dengan tepat sebesar 46,88 persen.

Pohon klasifikasi optimal yang terbentuk perlu divalidasi untuk mengetahui apakah pohon klasifikasi tersebut layak dan dapat digunakan untuk mengklasifikasi data baru. Tingkat keakuratan hasil klasifikasi pohon optimal yang dihasilkan dari data *testing* dapat dihitung berdasarkan Tabel 4.36.

Tabel 4.36 Klasifikasi Curah Hujan Data *Testing* pada Pohon Optimal Stasiun Pondok Betung Sebelum SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	0	0	0	0	0	0	0
2	1	5	0	0	0	83,33	1
3	0	1	0	0	0	0	1

Tabel 4.36 (Lanjutan) Klasifikasi Curah Hujan Data *Testing* pada Pohon Optimal Stasiun Pondok Betung Sebelum SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
4	0	0	0	0	0	0 %	0
5	0	0	0	0	0	0 %	0

Sehingga dapat dihitung besarnya ketepatan klasifikasi untuk data *testing* sebagai berikut:

$$1 - APER = \left(1 - \frac{2}{7}\right) \times 100\% = 71{,}43\%$$

Selanjutnya akan dibahas perhitungan ketepatan klasifikasi pohon optimal yang terbentuk setelah dilakukan proses SMOTE.

Tabel 4.37 Klasifikasi Curah Hujan Data *Learning* pada Pohon Optimal Stasiun Pondok Betung Setelah SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	61	6	4	1	0	85	11
2	9	187	63	33	7	63	112
3	4	39	60	7	4	53	54
4	3	16	12	89	0	74	31
5	0	3	0	0	125	98	3

Berdasarkan Tabel 4.37, kesalahan klasifikasi kelas pengamatan terjadi pada seluruh kelas. Ketepatan klasifikasi terbesar terjadi pada kelas 5 dengan persentase sebesar 98%. Dilanjutkan pada kelas 1 dengan persentase 85%. Sebanyak 112 pengamatan yang secara aktual termasuk kelas 2 (hujan ringan) namun salah diklasifikasikan sebagai sebagai kelas 1 (cerah berawan), 3 (hujan sedang), 4 (hujan lebat) dan 5 (hujan lebat sekali). Kesalahan klasifikasi juga terjadi pada kelas 3 (hujan sedang) dimana sebanyak 54 pengamatan berada pada kelas 1, 2, 4 dan 5. Selanjutnya sebanyak 31 pengamatan yang secara aktual masuk kelas 4 (hujan lebat), namun salah diklasifikasikan sebagai kelas 1 (cerah berawan), 2 (hujan ringan) dan 3 (hujan sedang), dan 4 (hujan lebat).

Menggunakan informasi pada Tabel 4.37, maka ketepatan klasifikasi data *learning* dapat dihitung sebagai berikut:

$$1 - APER = \left(1 - \frac{11 + 112 + 54 + 31 + 3}{733}\right) \times 100\% = 71{,}2\%$$

Hasil perhitungan ketepatan klasifikasi data *learning* sebesar 71,2 persen. Artinya pohon klasifikasi optimal mampu mengklasifikasikan pengamatan curah hujan kedalam kelas kategori hujan dengan tepat sebesar 71,2 persen.

Pohon klasifikasi optimal yang terbentuk perlu divalidasi untuk mengetahui apakah pohon klasifikasi tersebut layak dan dapat digunakan untuk mengklasifikasi data baru. Tingkat keakuratan hasil klasifikasi pohon optimal yang dihasilkan dari data *testing* dapat dihitung berdasarkan Tabel 4.38.

Tabel 4.38 Klasifikasi Curah Hujan pada Data *Testing* Pohon Optimal Stasiun Pondok Betung Setelah SMOTE

Actual	***Classified by Tree as***					**Ketepatan Klasifikasi (%)**	**Kesalahan Klasifikasi**
	1	**2**	**3**	**4**	**5**		
1	0	0	0	0	0	0	0
2	0	3	3	0	0	50	3
3	0	0	1	0	0	100	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0

Sehingga dapat dihitung besarnya ketepatan klasifikasi untuk data *testing* sebagai berikut:

$$1 - APER = \left(1 - \frac{3}{7}\right) \times 100\% = 57{,}14\%$$

Berikut adalah perbandingan hasil ketepatan klasifikasi pohon maksimal dengan pohon optimal yang ditunjukkan oleh Tabel 4.39.

Tabel 4.39 Perbandingan Ketepatan Klasifikasi Pohon Maksimal dan Pohon Optimal Stasiun Pondok Betung

	Pohon Klasifikasi	**Ketepatan Klasifikasi (%)**	
		Learning	***Testing* Data Baru**
Sebelum SMOTE	Pohon Maksimal	54,69	
	Pohon Optimal	46,88	71,43
Setelah SMOTE	Pohon Maksimal	71,50	
	Pohon Optimal	71,20	57,14

Berdasarkan Tabel 4.39, setelah dilakukan SMOTE terjadi peningkatan ketepatan klasifikasi pada data *learning*. Secara keseluruhan nilai ketepatan klasifikasi pohon maksimal lebih tinggi daripada pohon optimal, baik sebelum ataupun sesudah SMOTE. Hal ini dikarenakan pohon klasifikasi maksimal memiliki *node* yang paling banyak dengan melibatkan lebih banyak variabel prediktor sebagai pemilah *node* sehingga kemungkinan klasifikasi data dengan tepat akan cenderung lebih besar. Sedangkan pada ketepatan klasifikasi data *testing* terjadi penurunan setelah dilakukan SMOTE pada data pengamatan Pondok Betung. Penurunan ketepatan klasifikasi bisa jadi disebabkan oleh proses SMOTE yang kurang maksimal dalam mengatasi data *imbalance* untuk kelas lebih dari 2.

4.4 Perbandingan Hasil Ketepatan Klasifikasi Pohon pada Stasiun Pengamatan

Perbandingan hasil ketepatan klasifikasi dari analisis klasifikasi pohon untuk curah hujan pada 3 stasiun pengamatan yakni Citeko, Kemayoran dan Pondok Betung ditampilkan pada Tabel 4.40.

Tabel 4.40 Hasil Ketepatan Klasifikasi Seluruh Stasiun Pengamatan

Stasiun		**Pohon Klasifikasi**	**Ketepatan Klasifikasi (%)**	
			Learning	***Testing***
Citeko	Sebelum SMOTE	Pohon Maksimal	50,99	
		Pohon Optimal	7,95	100
	Setelah SMOTE	Pohon Maksimal	64,95	
		Pohon Optimal	63,58	28,57
Kemayoran	Sebelum SMOTE	Pohon Maksimal	51,19	
		Pohon Optimal	48,12	85,71
	Setelah SMOTE	Pohon Maksimal	67,20	
		Pohon Optimal	66,30	85,71
Pondok Betung	Sebelum SMOTE	Pohon Maksimal	54,69	
		Pohon Optimal	46,88	71,43
	Setelah SMOTE	Pohon Maksimal	71,50	
		Pohon Optimal	71,20	57,14

Dari Tabel 4.40, secara keseluruhan proses SMOTE memberikan kenaikan ketepatan klasifikasi cukup besar. Namun ketika dilakukan validasi dengan data baru sebagai data *testing*, proses SMOTE cenderung menurunkan ketepatan klasifikasi yang dihasilkan. Hal ini dikarenakan proses SMOTE kurang maksimal dalam mengatasi data *imbalance* jika diterapkan pada kasus dengan kelas lebih dari 2. Pada penelitian ini, pohon klasifikasi yang layak untuk mengklasifikasikan curah hujan pada ketiga stasiun pengamatan adalah model klasifikasi pohon optimal sebelum proses SMOTE.

BAB V

KESIMPULAN DAN SARAN

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis yang dilakukan, diperoleh kesimpulan sebagai berikut:

1. Hasil reduksi dimensi menggunakan metode PCA menghasilkan total komponen utama (PC) pada Stasiun Citeko sebanyak 42 komponen, Stasiun Kemayoran 36 komponen, dan Stasiun Pondok Betung sebanyak 37 komponen.
2. Hasil ketepatan klasifikasi curah hujan terbesar menggunakan data *testing* terletak pada Stasiun Pondok Betung. Berdasarkan hasil ketepatan klasifikasi data *testing* untu setiap stasiun pengamatan, maka pohon klasifikasi yang layak untuk klasifikasi curah hujan adalah model klasifikasi pohon optimal yang sebelum diproses menggunakan SMOTE.
3. Hasil ketepatan klasifikasi data *testing* sebelum proses SMOTE pada Stasiun Citeko, Kemayoran, dan Pondok Betung yakni 100%, 85,71% dan 71,43%. Setelah proses SMOTE, ketepatan klasifikasi ketiga stasiun pengamatan cenderung turun yakni 28,57%, 85,71% dan 57,14%.

5.2 Saran

Penelitian selanjutnya dapat menggunakan metode selain SMOTE untuk mengatasi kelas *imbalance* data terutama untuk kasus dengan kelas lebih dari 2 (*multi-class*). Sehingga diharapkan akan menghasilkan nilai akurasi yang lebih tinggi dan jumlah *node* terminal yang lebih sederhana untuk data curah hujan di Stasiun Pengamatan Citeko, Kemayoran dan Pondok Betung. Selain itu, saran untuk penelitian selanjutnya adalah

1. Diharapkan dapat menggunakan data *training* dan *testing* dengan periode yang lebih panjang.
2. Perlu dilakukan adanya imputasi data *missing* curah hujan, agar menghasilkan model yang lebih representatif terhadap data.

(Halaman ini sengaja dikosongkan)

DAFTAR PUSTAKA

DAFTAR PUSTAKA

[BMG] Badan Meteorologi dan Geofisika. (2008). *Model Atmosfer (CCAM) Conformal Cubic Atmospheric Model.* Padang: Pusat Penelitian dan Pengembangan.

[BMKG] Badan Meteorologi Klimatologi dan Geofisika. (2011). *Analisis Musim Hujan 2011/2012 dan Prakiraan Musim Kemarau 2012 Propinsi Banten dan DKI Jakarta.* Tangerang: BMKG-Pondok Betung.

[BMKG] Badan Meteorologi Klimatologi dan Geofisika. (2011). *Kajian dan Aplikasi Model CCAM (Conformal Cubic Atmospheric Model) untuk Prakiraan Cuaca Jangka Pendek Menggunakan MOS (Model Output Statistics).* Jakarta: Pusat Penelitian dan Pengembangan BMKG.

[BMKG] Badan Meteorologi Klimatologi dan Geofisika. (2015). *Analisis Musim Kemarau dan Prakiraan Musim Hujan 2015/2016.* Tangerang: BMKG Pondok Betung.

[BMKG] Badan Meteorologi, Klimatologi, dan Geofisika. (2016). *Badan Meteorologi, Klimatologi, dan Geofisika.* Retrieved Nopember 27, 2016, from Badan Meteorologi, Klimatologi, dan Geofisika Web Site: www.bmkg.go.id /iklim/prakiraan-musim.bmkg

Anuravega, A. (2012). *Post Processing Peramalan Unsur Cuaca dengan Model Output Statistics (MOS): Studi Perbandingan Antara Reduksi Dimensi Independent Component Analysis (ICA) dan Principal Component Analysis (PCA).* Tugas Akhir. Surabaya: Institut Teknologi Sepuluh Nopember.

Arfianto, A. D. (2008). *Aplikasi Model regresi Logistik Untuk Prakiraan Kejadian Hujan.* Skripsi. Bogor: Institut Pertanian Bogor.

Breiman, L. (1993). *Classification and Regression Trees.* New York: Chapman Hall.

Budiyanti, D. (2010). *Pemodelan Curah Hujan Bulanan di Kabupaten Ngawi dengan Metode Regresi Pohon Ber-*

dasarkan Indikator ENSO. Tugas Akhir. Surabaya: Institut Teknologi Sepuluh Nopember.

Hairani. (2016). *Metode Klasifikasi data Mining dan teknik Sampling SMOTE Menangani Class Imbalance untuk Segmentasi Customer pada Industri Perbankan.* Yogyakarta: Universitas Gajah Mada.

Idayati. (2014). *Reduksi Dimensi NWP dengan Transformasi Wavelet Diskrit dan PCA untuk Pra-pemrosesan Data Dalam Pemodelan Prakiraan Curah Hujan.* Tugas Akhir. Surabaya: Institut Teknologi Sepuluh Nopember.

Idowu & Rautanbach. (2009). *Model Output Statistics to Improve Severe Storm Prediction Over Western Sahel.* University of Pretoria, Geography, South Africa.

Johnson, R. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). New Jersey: Prentice Hall.

Lewis, R. (2000). *An Introduction to Classification and Regression Tree (CART) Analysis, Annual Meeting of the Society for Academic Emergency Medicine.* California: San Francisco.

Mosley, L. (2013). A Balanced Approach to The Multi-Class Imbalance Problem. *Graduate These and Dissertations.*

Nichols, M. (2008). *Model Output Statistics. Independent Research program.*

Ningrum, A. W. (2015). *Classification and Regression Tree untuk Pengklasifikasian Stastus Rumuh Tangga Terhadap Penyakit Malaria di Provinsi Papua Barat dengan Pra-Pemrosesan Synthetic Minority Oversampling Technique.* Surabaya: Institut Teknologi Sepuluh Nopember .

Paramita, P. S. (2010). *Klasifikasi Sifat Curah Hujan Berdasar-kan Indikator ENSO di Kabupaten Ngawi dengan Menggunakan Metode Klasifikasi Pohon.* Tugas Akhir. Surabaya: Institut Teknologi Sepuluh Nopember.

Prastuti, M. (2013). *Klasifikasi Kejadian Hujan Menggunakan Regresi Logistik Ordinal dan Principal Component Analysis Sebagai Pra-Pemrosesan Data Numerical*

Weather Prediction. Tugas Akhir. Surabaya: Institut Teknologi Sepuluh Nopember.

Safitri, R. (2012). *Model Output Statistics dengan Projection Pursuit Regression untuk Meramalkan Suhu Minimum, Suhu Maksimum, dan Kelembapan.* Tugas Akhir. Surabaya: Institut Teknologi Sepuluh Nopember.

Septiana, L. (2014). *Statistically Inspired Modification Of Partial Least Square Untuk Memprediksi Suhu Dan Kelembaban Dengan Pra-Pemrosesan Principal Component Analysis.* Tugas Akhir. Surabaya: Institut Teknologi Sepuluh Nopember.

Statsoft, I. (2003). *Classification and Regression Trees.* Retrieved September 16, 2016, from http://www.statsoft.com/textbook/stchaid.html

Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences* (2nd ed.). Boston: Elvesier.

Yusri, E. (2008). *Penerapan Metode Pohon Klasifikasi dengan Algoritm CART pada data Status Daerah Kabupaten di Indonesia.* Skripsi. Surabaya: Institut Pertanian Bogor.

(Halaman ini sengaja dikosongkan)

LAMPIRAN

LAMPIRAN

Lampiran 1: Rata-Rata dan Standar Deviasi Variabel NWP di Stasiun Citeko

Variabel (grid)	Rata-rata	Standar Deviasi	Variabel (grid)	Rata-rata	Standar Deviasi
dpsdt (1)	-0,3389	140,0363	rh1 (4)	77,4815	7,0263
dpsdt (2)	-0,5951	139,8776	rh1 (5)	78,5862	7,2046
dpsdt (3)	-0,8928	139,7818	rh1 (6)	80,9973	7,3009
dpsdt (4)	0,1177	139,2114	rh1 (7)	73,8705	7,9770
dpsdt (5)	-0,4233	139,2438	rh1 (8)	74,6719	7,1984
dpsdt (6)	-0,8064	139,3067	rh1 (9)	75,2541	6,5241
dpsdt (7)	0,7951	137,1379	rh2 (1)	78,3912	6,8198
dpsdt (8)	0,2516	137,7203	rh2 (2)	78,3912	6,9025
dpsdt (9)	0,2231	138,3827	rh2 (3)	78,5323	6,9978
mixr1 (1)	0,0163	0,0009	rh2 (4)	76,6487	7,0990
mixr1 (2)	0,0164	0,0009	rh2 (5)	76,8835	7,0861
mixr1 (3)	0,0165	0,0009	rh2 (6)	77,3636	7,0010
mixr1 (4)	0,0157	0,0010	rh2 (7)	75,3548	8,3453
mixr1 (5)	0,0158	0,0010	rh2 (8)	75,0973	7,6207
mixr1 (6)	0,0162	0,0010	rh2 (9)	74,2953	6,9868
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
psl (4)	1008,7800	3,1188	zg4 (4)	1046,8571	2,9286
psl (5)	1008,7649	3,1212	zg4 (5)	1025,4204	2,4842
psl (6)	1008,7469	3,1272	zg4 (6)	1000,9567	3,2033
psl (7)	1008,8994	3,1198	zg4 (7)	1219,4702	2,6066
psl (8)	1008,8201	3,1232	zg4 (8)	1127,7327	3,0055
psl (9)	1008,7335	3,1314	zg4 (9)	1034,7825	2,7201

Lampiran 2: Rata-Rata dan Standar Deviasi Variabel NWP di Stasiun Kemayoran

Variabel (grid)	Rata-rata	Standar Deviasi	Variabel (grid)	Rata-rata	Standar Deviasi
dpsdt (1)	5,1654	139,5781	rh1 (4)	87,1663	5,9920
dpsdt (2)	4,9038	139,5001	rh1 (5)	87,6510	5,9499
dpsdt (3)	4,6787	139,3140	rh1 (6)	88,1838	5,9824
dpsdt (4)	5,5304	138,8028	rh1 (7)	85,3895	5,9625
dpsdt (5)	5,2498	138,6288	rh1 (8)	85,6710	6,0032
dpsdt (6)	5,0130	138,4813	rh1 (9)	86,2285	6,0841
dpsdt (7)	5,9036	137,8741	rh2 (1)	80,9396	5,9435
dpsdt (8)	5,6913	137,4893	rh2 (2)	80,9255	5,8825
dpsdt (9)	5,3825	137,4275	rh2 (3)	80,9675	6,1038
mixr1 (1)	0,0172	0,0008	rh2 (4)	80,1052	6,1376
mixr1 (2)	0,0171	0,0008	rh2 (5)	80,1705	6,1275
mixr1 (3)	0,0172	0,0008	rh2 (6)	80,5424	6,1494
mixr1 (4)	0,0171	0,0008	rh2 (7)	79,9395	6,0503
mixr1 (5)	0,0171	0,0008	rh2 (8)	80,0769	5,9986
mixr1 (6)	0,0171	0,0008	rh2 (9)	80,3181	5,9919
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
psl (4)	1008,5981	3,0415	zg4 (4)	942,9767	2,3703
psl (5)	1008,5908	3,0437	zg4 (5)	934,1384	2,3476
psl (6)	1008,5940	3,0458	zg4 (6)	936,9378	2,3184
psl (7)	1008,6065	3,0517	zg4 (7)	967,0228	2,3727
psl (8)	1008,5999	3,0540	zg4 (8)	959,2685	2,3671
psl (9)	1008,5996	3,0567	zg4 (9)	960,2754	2,3429

Lampiran 3: Rata-Rata dan Standar Deviasi Variabel NWP di Stasiun Pondok Betung

Variabel (grid)	Rata-rata	Standar Deviasi	Variabel (grid)	Rata-rata	Standar Deviasi
dpsdt (1)	-2,2156	137,2701	rh1 (4)	85,0827	6,3454
dpsdt (2)	-2,4893	137,1285	rh1 (5)	85,7442	6,3115
dpsdt (3)	-2,7023	136.9161	rh1 (6)	85.,6796	6,2594
dpsdt (4)	-1,7419	136,3245	rh1 (7)	82,0430	6,6182
dpsdt (5)	-2,0461	136,2789	rh1 (8)	83,0881	6,7434
dpsdt (6)	-2,3059	136,1501	rh1 (9)	83,4578	6,7734
dpsdt (7)	-1,3278	135,5724	rh2 (1)	79,4385	7,0773
dpsdt (8)	-1,6404	135,4793	rh2 (2)	79,7654	7,0309
dpsdt (9)	-1,9197	135,3060	rh2 (3)	80,1587	6,8675
mixr1 (1)	0,0171	0,0009	rh2 (4)	79,2923	7,0238
mixr1 (2)	0,0171	0,0009	rh2 (5)	79,4824	7,0018
mixr1 (3)	0,0171	0,0009	rh2 (6)	79,6792	6,8710
mixr1 (4)	0,0168	0,0009	rh2 (7)	78,1337	7,2029
mixr1 (5)	0,0169	0,0009	rh2 (8)	78,2043	7,3430
mixr1 (6)	0,0168	0,0009	rh2 (9)	78,5066	7,2628
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
psl (4)	1008,6932	3,3291	zg4 (4)	959,5172	2,2904
psl (5)	1008,6921	3,3315	zg4 (5)	960,5434	2,2849
psl (6)	1008,6907	3,3332	zg4 (6)	964,2669	2,2701
psl (7)	1008,6979	3,3410	zg4 (7)	992,3534	2,3472
psl (8)	1008,6913	3,3443	zg4 (8)	989,2136	2,3605
psl (9)	1008,6881	3,3463	zg4 (9)	990,5146	2,3339

Lampiran 4: *Tree Sequence* Stasiun Pengamatan Citeko Sebelum SMOTE

```
=============
TREE SEQUENCE
=============
Dependent variable: HUJAN
   Terminal         Cross-Validated     Resubstitution   Complexity
 Tree Nodes          Relative Cost       Relative Cost    Parameter
 -----------------------------------------------------------------
    1     67        1.005 +/- 0.017               0.039        0.000
   29     11        1.042 +/- 0.021               0.329        0.014
   30     10        1.042 +/- 0.023               0.348        0.015
   31      9        1.040 +/- 0.023               0.370        0.017
   32      8        1.023 +/- 0.026               0.394        0.019
   33      6        1.027 +/- 0.024               0.449        0.022
   34      5        1.043 +/- 0.024               0.496        0.038
   35      4        1.052 +/- 0.024               0.546        0.040
   36**    3        0.953 +/- 0.060               0.601        0.044
   37      2        1.009 +/- 0.024               0.750        0.119
   38      1        1.000 +/- .161844E-03         1.000        0.200
```

Lampiran 5: *Tree Sequence* Stasiun Pengamatan Citeko Setelah SMOTE

```
=============
 TREE SEQUENCE
=============
Dependent variable: HUJAN
   Terminal         Cross-Validated     Resubstitution   Complexity
 Tree Nodes          Relative Cost       Relative Cost    Parameter
 -----------------------------------------------------------------
    1     79        0.386 +/- 0.021               0.053        0.000
    9**   42        0.384 +/- 0.021               0.118        0.003
   24     11        0.482 +/- 0.022               0.338        0.011
   25     10        0.483 +/- 0.022               0.357        0.015
   26      8        0.503 +/- 0.022               0.398        0.016
   27      7        0.518 +/- 0.022               0.426        0.023
   28      6        0.541 +/- 0.022               0.456        0.024
   29      5        0.549 +/- 0.022               0.494        0.030
   30      4        0.572 +/- 0.023               0.537        0.035
   31      3        0.661 +/- 0.020               0.595        0.046
   32      2        0.761 +/- 0.012               0.750        0.124
   33      1        1.000 +/- 0.000               1.000        0.200
```

Lampiran 6: Variabel Pemilah Pohon Maksimal Stasiun Citeko Sebelum SMOTE

<table>
<tr><th>Variabel</th><th>Skor</th><th></th><th>Variabel</th><th>Skor</th><th></th></tr>
<tr><td>PC2QGSCR</td><td>100</td><td>||||||||||||||||||||||||||||||||||||||||||</td><td>PC1OMG4</td><td>17,41</td><td>|||||||</td></tr>
<tr><td>PCDPSDT</td><td>72,82</td><td>||||||||||||||||||||||||||||||||||</td><td>PCPSL</td><td>15,14</td><td>|||||||</td></tr>
<tr><td>PCPBLH</td><td>72,05</td><td>||||||||||||||||||||||||||||||||</td><td>PCPS</td><td>15</td><td>|||||||</td></tr>
<tr><td>PCTSCRN</td><td>54,7</td><td>|||||||||||||||||||||||||</td><td>PC1OMG1</td><td>13,61</td><td>|||||||</td></tr>
<tr><td>PCTEMP1</td><td>53,1</td><td>|||||||||||||||||||||</td><td>PC2V2</td><td>13,21</td><td>|||||||</td></tr>
<tr><td>PC1ZG2</td><td>42,33</td><td>||||||||||||||||</td><td>PCRH1</td><td>12,44</td><td>||||||</td></tr>
<tr><td>PCTEMP2</td><td>37,87</td><td>|||||||||||||||</td><td>PC1QGSCR</td><td>12,14</td><td>|||||</td></tr>
<tr><td>PC1OMG2</td><td>36,96</td><td>||||||||||||||</td><td>PCU2</td><td>11,26</td><td>||||</td></tr>
<tr><td>PC1ZG4</td><td>34,59</td><td>||||||||||||||</td><td>PCMIXR2</td><td>11,03</td><td>||||</td></tr>
<tr><td>PC2OMG4</td><td>33,29</td><td>|||||||||||||</td><td>PC1V2</td><td>10,44</td><td>||||</td></tr>
<tr><td>PC1USTAR</td><td>32,68</td><td>|||||||||||</td><td>PC1V1</td><td>9,51</td><td>||||</td></tr>
<tr><td>PCV4</td><td>31,27</td><td>|||||||||||</td><td>PCTMAXSC</td><td>9,39</td><td>||||</td></tr>
<tr><td>PCU1</td><td>30,27</td><td>|||||||||||</td><td>PCRH4</td><td>8,72</td><td>||||</td></tr>
<tr><td>PCRND</td><td>28,37</td><td>||||||||||</td><td>PCMIXR1</td><td>8,54</td><td>|||</td></tr>
<tr><td>PC2OMG2</td><td>26,92</td><td>|||||||||</td><td>PCRH2</td><td>7,83</td><td>|||</td></tr>
<tr><td>PC2ZG4</td><td>26,16</td><td>|||||||||</td><td>PCTMINSC</td><td>7,77</td><td>||</td></tr>
<tr><td>PC2ZG2</td><td>25,69</td><td>||||||||</td><td>PCTPAN</td><td>6,46</td><td>||</td></tr>
<tr><td>PC2USTAR</td><td>21,82</td><td>||||||||</td><td>PC2OMG1</td><td>5,53</td><td>|</td></tr>
<tr><td>PCU4</td><td>20,06</td><td>||||||||</td><td>PCMIXR4</td><td>5,17</td><td>|</td></tr>
<tr><td>PCTEMP4</td><td>19,76</td><td>||||||||</td><td>PC2ZG1</td><td>2,67</td><td></td></tr>
<tr><td>PC2V1</td><td>18,71</td><td>|||||||</td><td>PC1ZG1</td><td>1,61</td><td></td></tr>
</table>

Lampiran 7: Variabel Pemilah Pohon Optimal Stasiun Citeko Sebelum SMOTE

<table>
<tr><th>Variabel</th><th>Skor</th><th></th><th>Variabel</th><th>Skor</th><th></th></tr>
<tr><td>PCPBLH</td><td>100</td><td>||||||||||||||||||||||||||||||||||||||||||</td><td>PCTEMP4</td><td>0</td><td></td></tr>
<tr><td>PCDPSDT</td><td>90,31</td><td>||||||||||||||||||||||||||||||||||||||</td><td>PCTMAXSC</td><td>0</td><td></td></tr>
<tr><td>PC2QGSCR</td><td>87,67</td><td>|||||||||||||||||||||||||||||||||||||</td><td>PCTMINSC</td><td>0</td><td></td></tr>
<tr><td>PC2OMG4</td><td>52,97</td><td>|||||||||||||||||||||||||</td><td>PCTPAN</td><td>0</td><td></td></tr>
<tr><td>PC2ZG4</td><td>48,42</td><td>|||||||||||||||||||||</td><td>PCTSCRN</td><td>0</td><td></td></tr>
<tr><td>PCRND</td><td>23,19</td><td>||||||||||||||||</td><td>PCU1</td><td>0</td><td></td></tr>
<tr><td>PCMIXR4</td><td>0</td><td></td><td>PCU2</td><td>0</td><td></td></tr>
<tr><td>PC1OMG1</td><td>0</td><td></td><td>PCRH1</td><td>0</td><td></td></tr>
<tr><td>PC2OMG2</td><td>0</td><td></td><td>PC1USTAR</td><td>0</td><td></td></tr>
<tr><td>PC1OMG4</td><td>0</td><td></td><td>PC2USTAR</td><td>0</td><td></td></tr>
<tr><td>PC1OMG2</td><td>0</td><td></td><td>PC1V1</td><td>0</td><td></td></tr>
<tr><td>PCPS</td><td>0</td><td></td><td>PC2V1</td><td>0</td><td></td></tr>
<tr><td>PCPSL</td><td>0</td><td></td><td>PC1V2</td><td>0</td><td></td></tr>
<tr><td>PC1QGSCR</td><td>0</td><td></td><td>PC2V2</td><td>0</td><td></td></tr>
<tr><td>PCMIXR1</td><td>0</td><td></td><td>PCV4</td><td>0</td><td></td></tr>
<tr><td>PCMIXR2</td><td>0</td><td></td><td>PC1ZG1</td><td>0</td><td></td></tr>
<tr><td>PCRH2</td><td>0</td><td></td><td>PC2ZG1</td><td>0</td><td></td></tr>
<tr><td>PCRH4</td><td>0</td><td></td><td>PC1ZG2</td><td>0</td><td></td></tr>
<tr><td>PC2OMG1</td><td>0</td><td></td><td>PC2ZG2</td><td>0</td><td></td></tr>
<tr><td>PCTEMP1</td><td>0</td><td></td><td>PC1ZG4</td><td>0</td><td></td></tr>
<tr><td>PCTEMP2</td><td>0</td><td></td><td>PCU4</td><td>0</td><td></td></tr>
</table>

Lampiran 8: Variabel Pemilah Pohon Maksimal Stasiun Citeko Setelah SMOTE

<table>
<tr><th>Variabel</th><th>Skor</th><th></th><th>Variabel</th><th>Skor</th><th></th></tr>
<tr><td>PCPBLH</td><td>100</td><td>||||||||||||||||||||||||||||||||||||||||||</td><td>PCMIXR4</td><td>21,14</td><td>||||||||</td></tr>
<tr><td>PC2QGSCR</td><td>97,43</td><td>|||||||||||||||||||||||||||||||||||||||||</td><td>PCMIXR1</td><td>20,98</td><td>||||||||</td></tr>
<tr><td>PCDPSDT</td><td>78,12</td><td>|||||||||||||||||||||||||||||||||</td><td>PC2V1</td><td>20,79</td><td>||||||||</td></tr>
<tr><td>PCRH2</td><td>68,89</td><td>|||||||||||||||||||||||||||||</td><td>PC2OMG2</td><td>20,58</td><td>||||||||</td></tr>
<tr><td>PCMIXR2</td><td>67,05</td><td>||||||||||||||||||||||||||||</td><td>PC2V2</td><td>20,28</td><td>||||||||</td></tr>
<tr><td>PC2ZG1</td><td>38,7</td><td>||||||||||||||||</td><td>PCTMINSC</td><td>19,82</td><td>||||||||</td></tr>
<tr><td>PC2ZG4</td><td>36,36</td><td>|||||||||||||||</td><td>PCTEMP1</td><td>18,7</td><td>|||||||</td></tr>
<tr><td>PCU1</td><td>31,76</td><td>|||||||||||||</td><td>PCTPAN</td><td>18,69</td><td>|||||||</td></tr>
<tr><td>PC2OMG4</td><td>30,26</td><td>||||||||||||</td><td>PC1QGSCR</td><td>17,98</td><td>|||||||</td></tr>
<tr><td>PCTMAXSC</td><td>28,12</td><td>|||||||||||</td><td>PC2OMG1</td><td>16,62</td><td>||||||</td></tr>
<tr><td>PCTSCRN</td><td>26,48</td><td>||||||||||</td><td>PC1ZG4</td><td>14,45</td><td>|||||</td></tr>
<tr><td>PC1ZG1</td><td>25,97</td><td>||||||||||</td><td>PCPS</td><td>13,46</td><td>|||||</td></tr>
<tr><td>PC1OMG1</td><td>24,79</td><td>||||||||||</td><td>PC2ZG2</td><td>12,88</td><td>|||||</td></tr>
<tr><td>PCU4</td><td>24,61</td><td>||||||||||</td><td>PCU2</td><td>12,05</td><td>||||</td></tr>
<tr><td>PCRH1</td><td>24,27</td><td>|||||||||</td><td>PC1ZG2</td><td>11,77</td><td>||||</td></tr>
<tr><td>PCRH4</td><td>23,88</td><td>|||||||||</td><td>PC1V1</td><td>10,72</td><td>||||</td></tr>
<tr><td>PCTEMP2</td><td>23,53</td><td>|||||||||</td><td>PC2USTAR</td><td>10,56</td><td>||||</td></tr>
<tr><td>PC1USTAR</td><td>22,82</td><td>|||||||||</td><td>PCPSL</td><td>9,27</td><td>|||</td></tr>
<tr><td>PCV4</td><td>22,29</td><td>|||||||||</td><td>PC1OMG2</td><td>7,63</td><td>||</td></tr>
<tr><td>PCRND</td><td>22,16</td><td>|||||||||</td><td>PCTEMP4</td><td>6,92</td><td>||</td></tr>
<tr><td>PC1OMG4</td><td>21,9</td><td>||||||||</td><td>PC1V2</td><td>4,38</td><td>|</td></tr>
</table>

Lampiran 9: Variabel Pemilah Pohon Optimal Stasiun Citeko Setelah SMOTE

Variabel	**Skor**		**Variabel**	**Skor**	
PCPBLH	100	||||||||||||||||||||||||||||||||||||||||||	PCMIXR1	18,6	|||||||
PC2QGSCR	96,91	|||||||||||||||||||||||||||||||||||||||||	PCRH1	18,18	|||||||
PCDPSDT	75,01	|||||||||||||||||||||||||||||||	PC1OMG4	18,02	|||||||
PCRH2	69,12	|||||||||||||||||||||||||||||	PC2V1	16,99	||||||
PCMIXR2	64,14	|||||||||||||||||||||||||||	PC2OMG2	16,47	||||||
PC2ZG1	39,65	||||||||||||||||	PCTPAN	16,32	||||||
PC2ZG4	34,03	||||||||||||||	PCMIXR4	16,29	||||||
PCU1	29,46	||||||||||||	PC1QGSCR	13,04	|||||
PC1ZG1	26,61	||||||||||	PCTEMP1	12,82	|||||
PCTMAXSC	25,64	||||||||||	PCPS	10,44	|||
PCTSCRN	22,99	|||||||||	PC2ZG2	10,33	|||
PC2OMG4	22,33	|||||||||	PC1ZG4	10,13	|||
PCRH4	21,68	||||||||	PC1ZG2	10,03	|||
PCU4	20,71	||||||||	PCU2	9,27	|||
PC1USTAR	20,42	||||||||	PC2OMG1	9,22	|||
PCV4	19,33	|||||||	PC2USTAR	7,53	||
PCTEMP2	19,04	|||||||	PCPSL	7,12	||
PC1OMG1	19,03	|||||||	PC1V1	6,98	||
PCRND	18,94	|||||||	PCTEMP4	5	|
PCTMINSC	18,93	|||||||	PC1V2	3,49	|
PC2V2	18,87	|||||||	PC1OMG2	2,23	

Lampiran 10: *Tree Sequence* Stasiun Pengamatan Kemayoran Sebelum SMOTE

```
=============
TREE SEQUENCE
=============
Dependent variable: HUJAN
```

Terminal Tree	Nodes	Cross-Validated Relative Cost	Resubstitution Relative Cost	Complexity Parameter
1	38	1.021 +/- 0.025	0.083	0.000
7**	26	0.977 +/- 0.030	0.128	0.005
15	12	1.077 +/- 0.027	0.306	0.019
16	11	1.080 +/- 0.027	0.332	0.021
17	8	1.088 +/- 0.027	0.415	0.022
18	7	1.099 +/- 0.026	0.444	0.024
19	6	1.094 +/- 0.027	0.475	0.025
20	5	1.139 +/- 0.026	0.517	0.033
21	4	1.069 +/- 0.062	0.561	0.035
22	3	1.073 +/- 0.062	0.615	0.043
23	2	1.158 +/- 0.058	0.750	0.108
24	1	1.000 +/- 0.000	1.000	0.200

Lampiran 11: *Tree Sequence* Stasiun Pengamatan Kemayoran Setelah SMOTE

```
=============
TREE SEQUENCE
=============
Dependent variable: HUJAN
```

Terminal Tree	*Nodes*	Cross-Validated Relative Cost	Resubstitution Relative Cost	Complexity Parameter
1	47	0.377 +/- 0.026	0.060	0.000
4**	40	0.375 +/- 0.026	0.073	0.002
20	11	0.468 +/- 0.026	0.318	0.012
21	10	0.469 +/- 0.026	0.336	0.015
22	9	0.470 +/- 0.026	0.358	0.018
23	8	0.489 +/- 0.026	0.381	0.018
24	7	0.527 +/- 0.026	0.409	0.023
25	5	0.559 +/- 0.027	0.469	0.024
26	4	0.569 +/- 0.026	0.511	0.033
27	3	0.606 +/- 0.018	0.583	0.058
28	2	0.754 +/- 0.004	0.750	0.134
29	1	1.000 +/- .603157E-04	1.000	0.200

Lampiran 12: Variabel Pemilah Pohon Maksimal Stasiun Kemayoran Sebelum SMOTE

Variabel	**Skor**		**Variabel**	**Skor**	
PC2ZG2	100		PCTMINSC	46.03	
PCRH1	95.43		PC1ZG2	45.11	
PCMIXR4	93.96		PCMIXR1	43.81	
PCTEMP2	91.32		PCPS	42.49	
PCZG4	84.86		PCUSTAR	39.04	
PCDPSDT	83.96		PC1ZG1	37.37	
PCMIXR2	76.21		PCU2	36.02	
PCRH2	69.3		PCV2	31.23	
PCQGSCRN	62.53		PCV1	23.88	
PCV4	58.27		PCU1	22.91	
PCTEMP4	56.11		PCTPAN	21.29	
PCPBLH	54.65		PC3ZG1	18.91	
PCTMAXSC	53.91		PCOMG4	18.48	
PCU4	52.85		PCTEMP1	15.34	
PCOMG2	51.74		PC2ZG1	14.21	
PCRH4	51.69		PC4ZG1	13.85	
PCPSL	47.68		PCTSCRN	11.8	
PCOMG1	46.76		PCRND	3.42	

Lampiran 13: Variabel Pemilah Pohon Optimal Stasiun Kemayoran Sebelum SMOTE

<table>
<tr><th>Variabel</th><th>Skor</th><th></th><th>Variabel</th><th>Skor</th><th></th></tr>
<tr><td>PC2ZG2</td><td>100</td><td>||||||||||||||||||||||||||||||||||||||||||</td><td>PCOMG1</td><td>41,95</td><td>|||||||||||||||||</td></tr>
<tr><td>PCTEMP2</td><td>95,57</td><td>||||||||||||||||||||||||||||||||||||||||</td><td>PCMIXR1</td><td>41,25</td><td>|||||||||||||||||</td></tr>
<tr><td>PCRH1</td><td>93,98</td><td>|||||||||||||||||||||||||||||||||||||||</td><td>PCU4</td><td>40,45</td><td>||||||||||||||||</td></tr>
<tr><td>PCMIXR4</td><td>92,38</td><td>|||||||||||||||||||||||||||||||||||||||</td><td>PC1ZG1</td><td>39,11</td><td>||||||||||||||||</td></tr>
<tr><td>PCZG4</td><td>88,81</td><td>|||||||||||||||||||||||||||||||||||||</td><td>PCPS</td><td>38,93</td><td>||||||||||||||||</td></tr>
<tr><td>PCDPSDT</td><td>80,88</td><td>||||||||||||||||||||||||||||||||||</td><td>PCUSTAR</td><td>38,9</td><td>||||||||||||||||</td></tr>
<tr><td>PCMIXR2</td><td>73,15</td><td>||||||||||||||||||||||||||||||</td><td>PCU2</td><td>34,83</td><td>||||||||||||||</td></tr>
<tr><td>PCRH2</td><td>65,66</td><td>|||||||||||||||||||||||||||</td><td>PCV2</td><td>30,79</td><td>||||||||||||</td></tr>
<tr><td>PCV4</td><td>60,98</td><td>|||||||||||||||||||||||||</td><td>PCV1</td><td>23,1</td><td>|||||||||</td></tr>
<tr><td>PCQGSCRN</td><td>60,83</td><td>|||||||||||||||||||||||||</td><td>PCTPAN</td><td>22,28</td><td>|||||||||</td></tr>
<tr><td>PCTEMP4</td><td>57,64</td><td>||||||||||||||||||||||||</td><td>PCU1</td><td>19,99</td><td>||||||||</td></tr>
<tr><td>PCRH4</td><td>54,09</td><td>||||||||||||||||||||||</td><td>PC3ZG1</td><td>19,79</td><td>||||||||</td></tr>
<tr><td>PCTMAXSC</td><td>53,73</td><td>||||||||||||||||||||||</td><td>PC2ZG1</td><td>14,88</td><td>|||||</td></tr>
<tr><td>PCOMG2</td><td>50,11</td><td>|||||||||||||||||||||</td><td>PC4ZG1</td><td>14,49</td><td>|||||</td></tr>
<tr><td>PCTMINSC</td><td>48,17</td><td>||||||||||||||||||||</td><td>PCOMG4</td><td>14,22</td><td>|||||</td></tr>
<tr><td>PC1ZG2</td><td>47,21</td><td>|||||||||||||||||||</td><td>PCTEMP1</td><td>12,8</td><td>|||||</td></tr>
<tr><td>PCPSL</td><td>45,09</td><td>||||||||||||||||||</td><td>PCTSCRN</td><td>10,53</td><td>||||</td></tr>
<tr><td>PCPBLH</td><td>42,88</td><td>|||||||||||||||||</td><td>PCRND</td><td>3,58</td><td>|</td></tr>
</table>

Lampiran 14: Variabel Pemilah Pohon Maksimal Stasiun Kemayoran Setelah SMOTE

<table>
<tr><th>Variabel</th><th>Skor</th><th></th><th>Variabel</th><th>Skor</th><th></th></tr>
<tr><td>PCTEMP2</td><td>100</td><td>||||||||||||||||||||||||||||||||||||||||||</td><td>PC1ZG1</td><td>27,9</td><td>|||||||||||</td></tr>
<tr><td>PCZG4</td><td>93,8</td><td>|||||||||||||||||||||||||||||||||||||||</td><td>PCU2</td><td>27,64</td><td>|||||||||||</td></tr>
<tr><td>PC1ZG2</td><td>84</td><td>|||||||||||||||||||||||||||||||||||</td><td>PCMIXR2</td><td>27,52</td><td>|||||||||||</td></tr>
<tr><td>PCU4</td><td>77,53</td><td>||||||||||||||||||||||||||||||||</td><td>PCOMG2</td><td>26,71</td><td>||||||||||</td></tr>
<tr><td>PCRND</td><td>71,54</td><td>||||||||||||||||||||||||||||||</td><td>PCV2</td><td>26,7</td><td>||||||||||</td></tr>
<tr><td>PCU1</td><td>64,24</td><td>|||||||||||||||||||||||||||</td><td>PCOMG4</td><td>23,55</td><td>|||||||||</td></tr>
<tr><td>PCPBLH</td><td>58,85</td><td>||||||||||||||||||||||||</td><td>PCPSL</td><td>23,31</td><td>|||||||||</td></tr>
<tr><td>PC2ZG2</td><td>50,91</td><td>|||||||||||||||||||||</td><td>PCV4</td><td>20,14</td><td>||||||||</td></tr>
<tr><td>PCOMG1</td><td>50,3</td><td>|||||||||||||||||||||</td><td>PCTMINSC</td><td>19,38</td><td>|||||||</td></tr>
<tr><td>PCTEMP4</td><td>50,06</td><td>|||||||||||||||||||||</td><td>PCMIXR1</td><td>18,88</td><td>|||||||</td></tr>
<tr><td>PCRH2</td><td>47,58</td><td>|||||||||||||||||||</td><td>PCUSTAR</td><td>18,06</td><td>|||||||</td></tr>
<tr><td>PCQGSCRN</td><td>45,87</td><td>|||||||||||||||||||</td><td>PCV1</td><td>17,85</td><td>|||||||</td></tr>
<tr><td>PCRH4</td><td>44,39</td><td>||||||||||||||||||</td><td>PCTPAN</td><td>16,82</td><td>||||||</td></tr>
<tr><td>PCMIXR4</td><td>35,33</td><td>||||||||||||||</td><td>PCDPSDT</td><td>13,95</td><td>|||||</td></tr>
<tr><td>PCTEMP1</td><td>32,95</td><td>|||||||||||||</td><td>PCRH1</td><td>11,65</td><td>||||</td></tr>
<tr><td>PCTMAXSC</td><td>31,47</td><td>|||||||||||||</td><td>PC2ZG1</td><td>9,43</td><td>|||</td></tr>
<tr><td>PC3ZG1</td><td>29,79</td><td>||||||||||||</td><td>PC4ZG1</td><td>9,43</td><td>|||</td></tr>
<tr><td>PCPS</td><td>29,43</td><td>||||||||||||</td><td>PCTSCRN</td><td>2,97</td><td></td></tr>
</table>

Lampiran 15: Variabel Pemilah Pohon Optimal Stasiun Kemayoran Setelah SMOTE

Variabel	**Skor**		**Variabel**	**Skor**	
PCTEMP2	100	||||||||||||||||||||||||||||||||||||||||||	PC1ZG1	27,48	|||||||||||
PCZG4	95,43	||||||||||||||||||||||||||||||||||||||||	PCU2	27,21	|||||||||||
PC1ZG2	85,46	||||||||||||||||||||||||||||||||||||	PCMIXR2	26,41	||||||||||
PCU4	78,88	|||||||||||||||||||||||||||||||||	PCOMG2	26,29	||||||||||
PCRND	71,88	||||||||||||||||||||||||||||||	PCV2	23,6	|||||||||
PCU1	64,45	|||||||||||||||||||||||||||	PCPSL	22,89	|||||||||
PCPBLH	59,87	|||||||||||||||||||||||||	PCOMG4	21,27	||||||||
PC2ZG2	51,8	|||||||||||||||||||||	PCV4	19,58	|||||||
PCOMG1	51,18	|||||||||||||||||||||	PCMIXR1	19,21	|||||||
PCTEMP4	50,93	|||||||||||||||||||||	PCTMINSC	17,98	|||||||
PCRH2	46,84	|||||||||||||||||||	PCV1	16,6	||||||
PCQGSCRN	45,91	|||||||||||||||||||	PCTPAN	16,36	||||||
PCRH4	42,16	|||||||||||||||||	PCUSTAR	15,09	|||||
PCTEMP1	33,52	|||||||||||||	PCDPSDT	14,19	|||||
PCMIXR4	32,48	|||||||||||||	PCRH1	11,85	||||
PCTMAXSC	30,7	||||||||||||	PC2ZG1	9,59	|||
PC3ZG1	30,31	||||||||||||	PC4ZG1	9,59	|||
PCPS	29,11	||||||||||||	PCTSCRN	3,02	

Lampiran 16: *Tree Sequence* Stasiun Pengamatan Pondok Betung Sebelum SMOTE

```
=============
 TREE SEQUENCE
=============
Dependent variable: HUJAN
   Terminal         Cross-Validated     Resubstitution   Complexity
 Tree Nodes          Relative Cost       Relative Cost    Parameter
 ------------------------------------------------------------------
    1     49        0.984 +/- 0.037              0.053        0.000
   11**   25        0.963 +/- 0.038              0.137        0.006
   15     18        0.997 +/- 0.038              0.204        0.009
   16     17        1.002 +/- 0.038              0.221        0.013
   17     15        1.004 +/- 0.038              0.255        0.014
   18     14        1.015 +/- 0.038              0.273        0.014
   19     13        1.014 +/- 0.038              0.299        0.021
   20      8        1.023 +/- 0.042              0.447        0.024
   21      7        1.057 +/- 0.042              0.491        0.035
   22      6        1.036 +/- 0.048              0.539        0.039
   23      2        1.011 +/- 0.031              0.813        0.055
   24      1        1.000 +/- 0.000              1.000        0.150
```

Lampiran 17: *Tree Sequence* Stasiun Pengamatan Pondok Betung Setelah SMOTE

```
=============
TREE SEQUENCE
=============
Dependent variable: HUJAN
   Terminal          Cross-Validated      Resubstitution    Complexity
 Tree Nodes           Relative Cost        Relative Cost     Parameter
 -----------------------------------------------------------------------
    1     74         0.335 +/- 0.021                0.053         0.000
    7**   56         0.321 +/- 0.020                0.077         0.002
   31     10         0.517 +/- 0.022                0.359         0.021
   32      9         0.522 +/- 0.022                0.388         0.023
   33      8         0.527 +/- 0.022                0.417         0.023
   34      7         0.550 +/- 0.022                0.448         0.025
   35      6         0.617 +/- 0.022                0.485         0.030
   36      5         0.654 +/- 0.021                0.525         0.031
   37      4         0.697 +/- 0.020                0.590         0.052
   38      3         0.704 +/- 0.020                0.660         0.056
   39      2         0.785 +/- 0.008                0.783         0.098
   40      1         1.000 +/- .572205E-04          1.000         0.173
```

Lampiran 18: Variabel Pemilah Pohon Maksimal Stasiun Pengamatan Pondok Betung Sebelum SMOTE

Variabel	**Skor**		**Variabel**	**Skor**	
PCUSTAR	100	|||||||||||||||||||||||||||||||||||||||||||||||||||	PCDPSDT	29,65	|||||||||||||
PCTSCRN	78,05	|||||||||||||||||||||||||||||||||||||||||||	PCOMG4	28,39	|||||||||||||
PCTPAN	76,86	|||||||||||||||||||||||||||||||||||||||||||	PCTEMP4	23,21	|||||||||
PCTEMP1	72,44	|||||||||||||||||||||||||||||||||||||||	PC2ZG2	23,2	|||||||||
PCTEMP2	64,2	|||||||||||||||||||||||||||||||	PCMIXR4	22,77	|||||||||
PC1ZG4	62,14	|||||||||||||||||||||||||||||||	PCOMG1	22,36	|||||||||
PCV1	56,96	|||||||||||||||||||||||||	PCV4	18,39	||||||||
PCU1	51,82	|||||||||||||||||||||||||	PCOMG2	17,89	||||||||
PC1ZG2	50,44	|||||||||||||||||||||||||	PC5ZG1	15,97	||||||
PCU2	49,56	|||||||||||||||||||||||	PC2ZG1	13,63	|||||
PCPSL	48,27	|||||||||||||||||||||||	PCRND	11,64	||||
PCRH1	48,17	|||||||||||||||||||||||	PC3ZG1	10,61	||||
PCPS	42,43	|||||||||||||||||||	PCMIXR2	9,49	|||
PCTMINSC	39,32	|||||||||||||||||	PCRH4	9,3	|||
PCU4	37,51	|||||||||||||||	PCRH2	9,16	|||
PCTMAXSC	35,88	|||||||||||||||	PCV2	5,94	||
PCQGSCRN	30,95	|||||||||||||	PC1ZG1	1,49	
PCPBLH	30,68	|||||||||||||	PC4ZG1	0,57	
PCMIXR1	29,65	|||||||||||||			

Lampiran 19: Variabel Pemilah Pohon Optimal Stasiun Pengamatan Pondok Betung Sebelum SMOTE

<table>
<tr><th>Variabel</th><th>Skor</th><th></th><th>Variabel</th><th>Skor</th><th></th></tr>
<tr><td>PCUSTAR</td><td>100</td><td>||||||||||||||||||||||||||||||||||||||||||</td><td>PCTEMP4</td><td>20,84</td><td>||||||||</td></tr>
<tr><td>PCTSCRN</td><td>82,64</td><td>|||||||||||||||||||||||||||||||||||</td><td>PCMIXR1</td><td>19,08</td><td>|||||||</td></tr>
<tr><td>PCTPAN</td><td>77,98</td><td>|||||||||||||||||||||||||||||||||</td><td>PC5ZG1</td><td>16,9</td><td>||||||</td></tr>
<tr><td>PCTEMP1</td><td>72,85</td><td>||||||||||||||||||||||||||||||</td><td>PCQGSCRN</td><td>15,49</td><td>||||||</td></tr>
<tr><td>PCTEMP2</td><td>62,25</td><td>||||||||||||||||||||||||||</td><td>PCV4</td><td>13,19</td><td>|||||</td></tr>
<tr><td>PCV1</td><td>60,31</td><td>|||||||||||||||||||||||||</td><td>PC2ZG2</td><td>12,89</td><td>|||||</td></tr>
<tr><td>PC1ZG4</td><td>59,53</td><td>|||||||||||||||||||||||||</td><td>PCMIXR4</td><td>12,37</td><td>||||</td></tr>
<tr><td>PCPSL</td><td>49,48</td><td>||||||||||||||||||||</td><td>PC3ZG1</td><td>11,23</td><td>||||</td></tr>
<tr><td>PC1ZG2</td><td>46,41</td><td>|||||||||||||||||||</td><td>PCRND</td><td>11,14</td><td>||||</td></tr>
<tr><td>PCU2</td><td>44,24</td><td>||||||||||||||||||</td><td>PCOMG1</td><td>9,98</td><td>|||</td></tr>
<tr><td>PCU1</td><td>44,23</td><td>||||||||||||||||||</td><td>PCOMG2</td><td>9,5</td><td>|||</td></tr>
<tr><td>PCRH1</td><td>44,17</td><td>||||||||||||||||||</td><td>PC2ZG1</td><td>9,03</td><td>|||</td></tr>
<tr><td>PCPS</td><td>43,29</td><td>||||||||||||||||||</td><td>PCRH4</td><td>7,09</td><td>||</td></tr>
<tr><td>PCTMINSC</td><td>35,75</td><td>||||||||||||||</td><td>PCRH2</td><td>6,57</td><td>||</td></tr>
<tr><td>PCU4</td><td>32,5</td><td>|||||||||||||</td><td>PCV2</td><td>6,29</td><td>||</td></tr>
<tr><td>PCPBLH</td><td>28,72</td><td>|||||||||||</td><td>PCMIXR2</td><td>5,71</td><td>|</td></tr>
<tr><td>PCTMAXSC</td><td>27,95</td><td>|||||||||||</td><td>PC1ZG1</td><td></td><td></td></tr>
<tr><td>PCDPSDT</td><td>27,09</td><td>|||||||||||</td><td>PC4ZG1</td><td></td><td></td></tr>
<tr><td>PCOMG4</td><td>22,81</td><td>|||||||||</td><td></td><td></td><td></td></tr>
</table>

Lampiran 20: Variabel Pemilah Pohon Maksimal Stasiun Pengamatan Pondok Betung Setelah SMOTE

Variabel	**Skor**		**Variabel**	**Skor**	
PCMIXR2	100		PC1ZG2	42,4	
PCPS	87,5		PCTMINSC	42,03	
PCTPAN	76,29		PCTEMP2	40,21	
PCRH4	71,72		PCU2	38,8	
PCPSL	71,27		PCTMAXSC	37,83	
PCV4	66,49		PCUSTAR	36,48	
PCV1	64,91		PC4ZG1	34,87	
PC1ZG4	64,08		PCOMG2	34,81	
PCRH1	63,35		PCOMG1	32,02	
PCPBLH	62,4		PC5ZG1	30,76	
PCQGSCRN	57,78		PCU4	30,65	
PC2ZG2	57,61		PCDPSDT	30,49	
PCTEMP4	57,6		PC1ZG1	29,08	
PCMIXR1	55,35		PC3ZG1	26,67	
PCTSCRN	52,63		PCOMG4	23,02	
PCRH2	52,01		PC2ZG1	19,55	
PCTEMP1	51,22		PCRND	18,42	
PCV2	50,12		PCU1	14,33	
PCMIXR4	45,81				

Lampiran 21: Variabel Pemilah Pohon Optimal Stasiun Pengamatan Pondok Betung Setelah SMOTE

Variabel	**Skor**		**Variabel**	**Skor**	
PCMIXR2	100		PCMIXR4	42,66	
PCPS	85,22		PCTMINSC	39,87	
PCTPAN	74,66		PCU2	39,48	
PCPSL	69,67		PCTEMP2	39,39	
PCRH4	67,87		PCUSTAR	35,69	
PCV1	65,21		PC4ZG1	34,65	
PCV4	64,69		PCTMAXSC	31,59	
PC1ZG4	62,95		PC5ZG1	31,29	
PCRH1	59,2		PCU4	31,17	
PCPBLH	58,19		PCOMG2	30,53	
PC2ZG2	57,18		PCOMG1	29,35	
PCQGSCRN	56,92		PC1ZG1	28,76	
PCTSCRN	53,54		PC3ZG1	26,31	
PCTEMP4	53,49		PCDPSDT	25,98	
PCMIXR1	52,52		PCOMG4	19,67	
PCTEMP1	52,1		PC2ZG1	19,07	
PCRH2	50,1		PCRND	17,11	
PCV2	49,14		PCU1	12,96	
PC1ZG2	43,14				

(Halaman ini sengaja dikosongkan)

BIODATA PENULIS

BIODATA PENULIS

Penulis Tugas Akhir ini bernama lengkap Ulul Azmi, lahir di Jombang, pada tanggal 10 Juni 1993. Penulis merupakan anak kedua dari pasangan Bapak Sutrisno dan Ibu Alifah. Riwayat pendidikan penulis dimulai dari TK Aisiyah Bustanul Athfal Denpasar, SD Muhammadiyah 3 Denpasar. Kemudian setelah lulus SD, penulis memilih hijrah ke Jombang untuk *mondok* dan sekolah di SMP A. Wahid Hasyim Tebuireng Jombang, dilanjutkan ke SMA A. Wahid Hasyim Tebuireng Jombang. Terakhir penulis menempuh pendidikan di Institut Teknologi Sepuluh Nopember Surabaya jurusan Statistika pada tahun 2011 dengan NRP 1311100702. Selama di ITS penulis terlibat aktif dalam organisasi CSS MoRA ITS sebagai staf departemen humas pada tahun 2012 dan berkesempatan menjadi sekretaris-bendahara humas pada tahun 2013. Pada masa kuliah penulis juga pernah melakukan kerja praktek di Kimia Farma *Trade* & *Distribution* Sidoarjo sebagai upaya pengaplikasian ilmu statistika di dunia nyata. Untuk menyelesaikan pendidikan di jenjang sarjana ini, penulis mengambil Tugas Akhir dengan tema klasifikasi-*data mining* dengan judul **"Prediksi Curah Hujan Melalui *Model Output Statistics* Menggunakan *Classification and Regression Trees* dengan *Pre-processing Principal Component Analysis*"**. Jika pembaca ingin memberikan kritik dan saran serta ingin berdiskusi lebih lanjut, dapat menghubungi melalui alamat email: azmiarsyd@gmail.com

(Halaman ini sengaja dikosongkan)

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini, mahasiswa Jurusan Statistika FMIPA ITS:

Nama : Ulul Azmi
NRP : 1311 100 702

menyatakan bahwa data yang digunakan dalam Tugas Akhir/Thesis ini merupakan data sekunder yang diambil dari penelitian / buku/ Tugas Akhir/ Thesis/ publikasi lainnya yaitu:

Sumber : 1. Badan Meteorologi Klimatologi dan Geofisika
2. *NWP Arpeg Tropic Products* Meteo Franc

Keterangan : 1. Data Curah Hujan harian Jabodetabek periode 01 Januari 2009 – 31 Desember 2010
2. *Output* NWP periode 01 Januari 2009 - 31 Desember 2010

Surat Pernyataan ini dibuat dengan sebenarnya. Apabila terdapat pemalsuan data maka saya siap menerima sanksi sesuai aturan yang berlaku.

Mengetahui
Pembimbing Tugas Akhir

(Dr. Sutikno, S.Si, M.Si)
NIP. 19710313 199702 1 001

Surabaya, Januari 2017

(Ulul Azmi)
NRP. 1311 100 702

*(coret yang tidak perlu)

(Halaman ini sengaja dikosongkan)